MÁSTER OFICIAL EN EMPRESA Y TECNOLOGÍAS DE LA INFORMACIÓN

CURSO 2018-2019

TRABAJO FIN DE MÁSTER

Herramientas de Inteligencia de Negocio para la explotación y análisis de fuentes Open Data

(Business Intelligence tools for exploitation and analysis of Open Data sources)

Autor: Adrián Monge Sainz

Dirigido por: Rocío Rocha Blanco

Febrero 2019

Resumen

Las iniciativas Open Data de las Administraciones Públicas deben ser valoradas no solo por la cantidad de información que proporcionan, sino también por la implementación de mecanismos y uso de tecnologías que favorezcan su accesibilidad y reutilización. En este trabajo se hace una revisión del estado del arte de la corriente Open Data en España, haciendo énfasis en el portal de datos abiertos de la ciudad de Santander sobre el cual se aplica la métrica Meloda con el objetivo de hacer una evaluación exhaustiva del portal y señalar pautas que favorezcan la reutilización de los datos.

Para poner en valor dicha información es necesario el uso de diferentes herramientas que permita la extracción, la transformación y el análisis de grandes cantidades de datos, en ocasiones procedentes de diferentes fuentes y disponibles en múltiples formatos. En este sentido, se analizan diversas técnicas y metodologías de análisis de datos, así como las diferentes fases de vida de un proyecto de este tipo, finalizando el el trabajo con la realización de un caso práctico en el que se emplean herramientas de Business Intelligence y Minería de Datos para la generación de un modelo de datos que permita relacionar la información obtenida y obtener resultados a partir de informes. Para ello, se utilizarán datos procedentes de conjuntos de datos de mediciones de tráfico tomadas por sensores en tiempo real complementados con información georreferenciada que permite ubicar los sensores sobre un mapa a través de la tecnología GIS.

Abstract

Open Data initiatives of Public Administrations should be valued not only for the amount of information that has been provided, but also for the implementation of mechanisms and use of technologies that favor their accessibility and reuse. In this project, we have a review of the state of the art of the Open Data stream in Spain, with an emphasis on the data portal of the city of Santander, on which the Meloda metric is applied in order to make a thorough evaluation of the portal and point out that favor the reuse of the data.

To value such information, it is necessary to use different tools that allow the extraction, transformation and analysis of big amounts of data. In this sense, we analyse various techniques and methodologies of data analysis, as well as the different phases of life in a project of this type, ending the work with the realization of a practical case in which business intelligence tools are used and Data mining for the generation of a data model that allows to relate the information to obtain and obtain results from reports. For this, you can use the contact data of the data in real time complemented with the georeferenced information that allows you to locate the sensors on a map through GIS technology.

ÍNDICE DE CONTENIDOS

CAPÍTULO 1. INTRODUCCIÓN

El interés de las administraciones públicas por mejorar la calidad de vida de los ciudadanos residentes en las principales ciudades es cada vez mayor, y en este sentido el uso de las TIC juega un papel importante. Según el Estudio y Guía Metodológica sobre Ciudades Inteligentes elaborado por Deloitte en 2015, fue en el año 2010 con la publicación de la Estrategia Europa 2020 cuando se comenzara a priorizar un uso más eficiente de los recursos, en busca de una economía más verde y competitiva. Para alcanzar este y otros objetivos del plan, la Comisión Europea propuso la Agenda Digital para Europa, la cual sería adoptada por el gobierno español en el año 2013, estableciendo como una de las principales estrategias la elaboración de un Plan Nacional de Ciudades Inteligentes (2015) con el que se pretende ayudar a los diferentes organismos locales en su transformación digital hacia la ciudad inteligente que aproveche el uso de las nuevas tecnologías para monitorizar los diferentes espacios de la ciudad en búsqueda del ahorro energético, la reducción de emisiones contaminantes, el control del tráfico en la ciudad o la resolución de incidentes notificados por los ciudadanos entre otros fines.

1.1. MOTIVACIÓN Y JUSTIFICACIÓN

Visto el grado de importancia que las Administraciones Públicas (AAPP) tanto a nivel nacional como europeo están dando a la apertura de datos en sus proyectos de transformación digital, permitiendo al ciudadano acceder a grandes cantidades de datos públicos, así como su reutilización para la creación de aplicaciones o para su uso en diferentes formatos, la principal motivación de este trabajo es comprender el funcionamiento de los portales abiertos (*Open Data*), su estructura y las tecnologías que emplea para la publicación de la información.

Además de los beneficios sociales para el ciudadano que pueden tener los portales *Open Data* (transferencia de información, creación de apps, informes públicos para la ciudadanía…), destacan los beneficios económicos que estos pueden generar a través de empresas del sector infomediario, además de los beneficios que reportan a la imagen de las propias administraciones públicas que son percibidas como organismos más transparentes y cercanos al ciudadano.

Para poner en valor estos datos, en este trabajo surge la necesidad de estudiar las diferentes formas de valorar la calidad y relevancia de la información, además de seleccionar diferentes herramientas para su visualización y procesado, prestando especial atención a cómo pueden servir de ayuda dichas herramientas para el análisis de datos procedentes de fuentes en tiempo real.

1.2. OBJETIVOS GENERALES Y ESPECÍFICOS

Los objetivos generales de este trabajo se centran en profundizar en el uso de datos abiertos que se encuentran disponibles para el ciudadano en los diferentes portales *Open Data* proporcionados por organismos públicos y privados, el estudio y aprendizaje de técnicas y métricas para la evaluación de la calidad de la información en portales de datos y así como en el uso de herramientas de inteligencia de negocio y minería de datos con el objetivo de explotar dichos datos y convertirlos en conocimiento.

En cuanto a los objetivos específicos del trabajo se plantean los siguientes:

- Aportar una visión general del estado del arte de la corriente *Open Data* en España, así como de otras iniciativas de interés en la que se fomenta la apertura de los datos.

- Adquirir las competencias para el análisis de la estructura de los portales *Open Data*, consulta y acceso a los datos y valoración de la calidad de éstos.

- Exponer los fundamentos teóricos de *Business Intelligence* y Minería de datos, así como la evolución en las metodologías utilizadas en proyectos de dichas temáticas, con el fin de adquirir una visión general sobre las diferentes herramientas que serán utilizadas posteriormente para el análisis de los datos procedentes de portales de datos abiertos.

- Investigar y comparar las principales metodologías para llevar a cabo un proyecto de Minería de Datos en el ámbito de la Inteligencia de Negocio.

- Aplicar algunas de las estrategias y herramientas de *Business Intelligence* y *Data Mining* definidas en el marco teórico del documento para la realización de un caso práctico en el que se analizarán datos procedentes de fuentes de portales *Open Data*.

CAPÍTULO 2. INICIATIVAS DE DATOS ABIERTOS

En ámbitos no empresariales tales como el sector público, educativo etc. cobran una gran importancia las iniciativas de datos abiertos. Además de las conocidas iniciativas *Open Data* o RISP existen multitud de portales y comunidades de *Data Scientists* que ponen a disposición del usuario de forma gratuita *datasets* con información masiva organizada en diferentes categorías.

En este capítulo, se hará mención y se estudiará brevemente la situación actual de algunas de las iniciativas y portales de datos más utilizados hoy en día por los usuarios con el objetivo de generar conocimiento a partir del desarrollo de aplicaciones o del mero análisis de datos a partir de las técnicas comentadas en el anterior capítulo.

2.1. LA CORRIENTE OPEN DATA

La iniciativa *Open Data* es una práctica cuyo fin es que ciertos datos de los que disponen Administraciones Públicas (AAPP) y en ocasiones organizaciones privadas sean accesibles al ciudadano de forma libre y sin ningún tipo de restricción legal tales como pueden ser los derechos de autor o patentes. Esta iniciativa busca promover la transparencia, la igualdad de condiciones en lo que se refiere al acceso a la información por parte del ciudadano y la colaboración entre administraciones y empresas o particulares que tienen la oportunidad de crear aplicaciones a partir de ciertos datos las cuales pueden contribuir al beneficio común.

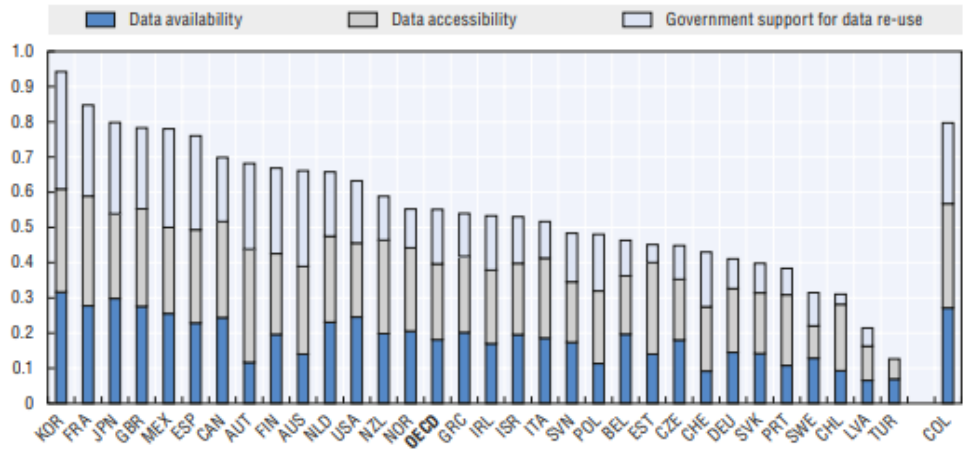
En este sentido, uno de los principales enfoques de la filosofía de datos abiertos es la Reutilización de la Información en el Sector Público (RISP) que supone el acceso a múltiples fuentes de datos de interés general facilitados por las AAPP en distintos ámbitos (geográfico, demográfico, social, económico…). De este modo, el principal objetivo de RISP según el portal de Gobierno abierto de Navarra es conseguir la distribución y reutilización de dichos datos para lograr la colaboración entre empresas, ciudadanos y administraciones públicas con el fin de conseguir un gobierno más abierto, promover la creación de valor y fomentar la interoperabilidad entre diferentes organismos.

Actualmente existen varios rankings e informes que miden la calidad de los portales de datos abiertos en todo el mundo. Uno de los más actuales es el *Our Data Index* de la OCDE, que a través del informe *Government at a Glan*ce publicado en 2017 mide la calidad de los portales de datos abiertos de los 35 países que en ese año constituían la la Organización para la Cooperación y el Desarrollo Económico. Los resultados de este estudio están basados en 140 indicadores que evalúan tres características básicas: reutilización de los datos y apoyo gubernamental para dicha reutilización, disponibilidad y accesibilidad.

En este ranking liderado por Corea, España se sitúa en el sexto lugar por detrás de países europeos como Francia o Gran Bretaña. Comparando estos resultados con los del anterior informe, publicado en 2015, España se mantiene constante en el ranking con una valoración de 0,77 puntos en una escala entre 0 y 1 en comparación con la del pasado informe en la que se situaba en los 0,78 puntos. Por su parte, la media de los países de los países OCDE desciende hasta los 0,55 puntos en comparación con los resultados emitidos en 2015 donde se situaba en los 0,58 de media. A pesar del descenso general experimentado, el último informe destaca los grandes avances que han realizado estos países en la apertura de sus datos al ciudadano. Sin embargo, insta a los gobiernos un mayor esfuerzo por facilitar la reutilización de dichos datos ya que,

como se puede observar en el **Figura 2.1**, la aportación de dicho indicador a la puntuación total de la mayoría de los países es la más baja en comparación con los criterios de disponibilidad y accesibilidad.

Figura 2.1: Ranking mundial de iniciativas Open Data. Fuente: Our Data Index 2017



2.1.1. Situación en España

En el caso de España, las iniciativas de apertura de datos por parte de las administraciones públicas nacen en 2005 con el inicio del *Plan Avanza*, desde el 2009 gestionado desde la *Proyecto Aporta*, la cual se sustenta en tres objetivos: "Fomentar una cultura favorable a la apertura de datos públicos, facilitar que las administraciones acometan dicha apertura e impulsar el mercado de la reutilización de la información pública."[1] Dos de los principales logros de dicho proyecto han sido, por un lado, la creación del portal datos.gob.es, el cual recopila todos los catálogos de datos publicados por las diferentes instituciones a nivel estatal, regional y ayuntamientos y, por otro lado, la publicación de una guía con consejos y buenas prácticas para facilitar la creación de portales *Open Data*. En la actualidad, el proyecto ha pasado a denominarse con el mismo nombre que el propio portal.

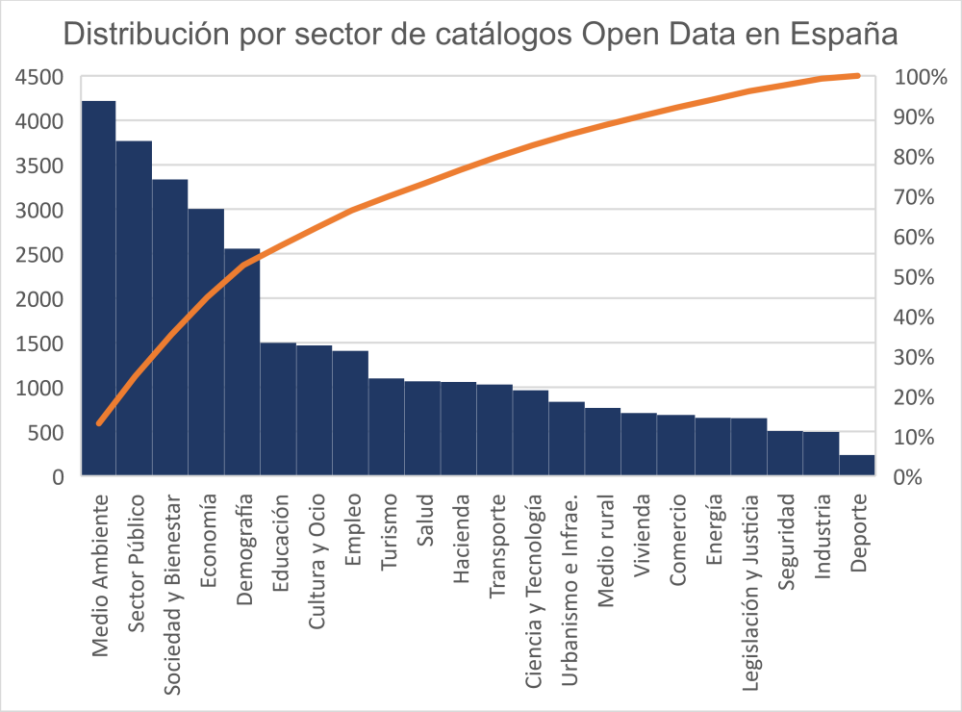
En cuanto a las iniciativas en activo actualmente, la Fundación Centro Tecnológico de la Información y la Comunicación (CTIC) dispone de un mapa actualizado con las iniciativas *Open Data* llevadas a cabo a nivel estatal, regional y local. En el mapa de la **Figura 2.2** se ubican los 46 catálogos disponibles en agosto de 2018.

[1] http://datos.gob.es/es/documentacion/proyecto-aporta

Figura 2.2: Iniciativas Open Data en España[2]. Elaboración propia a partir del mapa de la Fundación CTIC [Consultado el 27/08/2018]

Por otro lado, en cuanto a la reutilización de información del sector público en España, destaca la presencia de catálogos relacionados con temáticas sobre medio ambiente, sector público, sociedad y bienestar, datos económicos y datos demográficos, representando estas cinco categorías alrededor del 60% de los catálogos totales disponibles en los diferentes portales *Open Data* españoles. En la **Figura 2.3** se representa la distribución de catálogos de datos abiertos clasificados por sector en orden descendente, así como la línea de Pareto que representa el porcentaje acumulado de datos representados en la gráfica.

Figura 2.3: Distribución por sector de catálogos Open Data en España. Elaboración propia a partir de la clasificación por categorías de datos.gob.es [Consultado el 29/10/2018]



[2] Excluidos los catálogos pertenecientes al sector público español por considerarse dentro del portal datos.gob.es

2.1.2. Marco tecnológico

En esta sección se explicarán las estructuras más comunes de los portales *Open Data*, haciendo énfasis en los tipos de datos que estos pueden ofrecer y en los términos propios utilizados en el ámbito de los datos abiertos. En la mayor parte de los portales se ofrecen varias secciones diferenciadas haciendo referencia principalmente a los catálogos de datos disponibles, a la realización de consultas SPARQL y al uso de APIs que permiten a los usuarios colaborar mediante el diseño de aplicaciones específicas a partir de los datos y herramientas facilitados en el portal. Además, en el caso de los portales pertenecientes a las ciudades catalogadas como *Smart cities* es habitual que se publiquen *datasets* actualizados en tiempo real con datos procedentes de sensores situados en puntos estratégicos de la ciudad.

2.1.2.1. Datos vinculados

Cada vez es más habitual la publicación de los conjuntos de datos lleva asociada información que aporta un valor añadido a los datos (metadatos). No obstante, gran parte de la información disponible en la web no se encuentra vinculada, es decir, las máquinas no pueden interpretar de forma automática las relaciones presentes. Para ir un paso más allá, surge el concepto de datos enlazados o *Linked Data*. Este método de publicación avanzado utiliza estándares web como HTTP, RDF o URI con el objetivo de dotar a los datos de mayor usabilidad. A continuación, se definen los conceptos asociados a este método:

- *HTTP (Hypertext Transfer Protocol)*: protocolo de comunicaciones que permite la transferencia de información en la web. En los modelos de datos enlazados se utilizan URIs HTTP comprensibles por humanos y máquinas.

- *URI (Uniform Resource Identifier)*: conjunto de caracteres que identifica al conjunto de datos a través de la siguiente estructura: esquema (protocolo utilizado: http, ftp…), autoridad (identifica el sitio web), ruta (identifica el recurso de forma jerárquica), query string (se trata de un par clase=valor precedido por un signo “?” cuyo objetivo es solicitar información de la base de datos para mostrarla en la aplicación web.) y fragmento (apunta a una parte concreta de la página y se identifica a través del símbolo “#”).

- *RDF (Resource Description Framework)*: es un conjunto de especificaciones de W3C utilizados como modelo de datos para metadatos. Este modelo representa las relaciones a través de triplas (sujeto, predicado, objeto) de forma que el sujeto es el recurso que se quiere definir, el predicado es la relación y el objeto es otro recurso con el que se relaciona el primero. A través de este estándar se aporta información útil para cada URI.

2.1.2.2. Consultas SPARQL

SPARQL es un lenguaje de consultas utilizado en repositorios de datos abiertos que cumplen con el estándar RDF. Dado que los datos que siguen el estándar RDF constituyen bases de datos orientadas a grafos, es necesario un lenguaje que permita acceder a dichos datos y obtenerlos en forma de tablas, de forma que sean manejables por herramientas de análisis de datos y visualización (Excel, Power BI etc.). De este modo, el lenguaje SPARQL se encarga de interpretar este tipo de bases de datos

2.1.2.3. APIs y aplicaciones

Las interfaces de programación de aplicaciones constituyen una herramienta que incluye las principales funcionalidades del portal *Open Data* en cuestión, de forma que puede ser utilizada para el despliegue de una nueva aplicación basada en los contenidos del portal. La principal finalidad de las APIs en este caso no es proporcionar una interfaz de usuario con funciones predefinidas como punto de partida para el desarrollador, sino que suelen utilizarse para aportar rutas que sirvan como punto de acceso a la información del portal, habitualmente a través de consultas SPARQL. De este modo, las aplicaciones puedan disponer de los datos de forma actualizada a través de la realización recurrente de consultas. como alternativa al acceso a través de la aplicación web.

Otra tecnología muy extendida en los portales de datos abiertos es la denominada API REST. Un servicio web REST propone una solución más sencilla de manipulación de datos que la de las APIs tradicionales basadas en el protocolo SOAP (Simple Open Access Protocol). Según (BBVA, 2016), "*REST es cualquier interfaz entre sistemas que use HTTP para obtener datos o generar operaciones sobre esos datos en todos los formatos posibles, como XML y JSON.*"

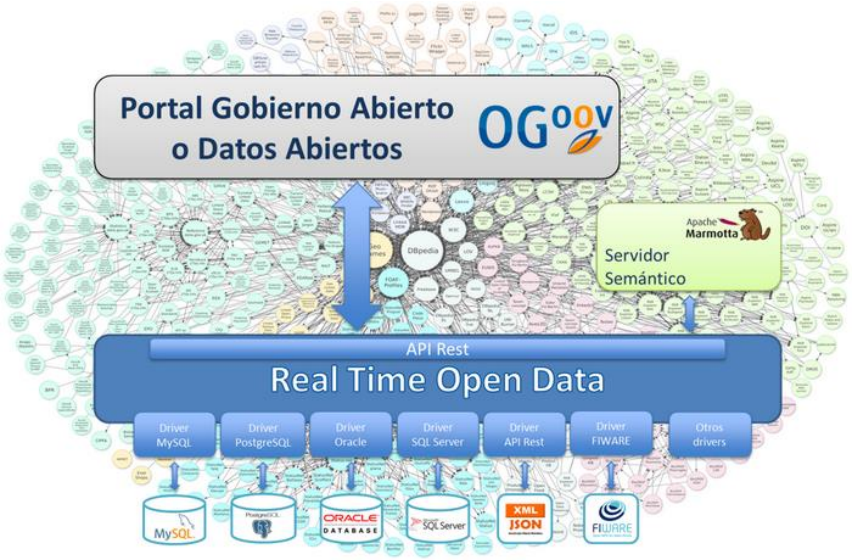
2.1.2.4. Datos en tiempo real

Uno de los grandes retos de las administraciones públicas es gestionar las grandes cantidades de datos procedentes de las medidas de sensores y otros sistemas situados en las ciudades inteligentes. Es tal la cantidad de datos que dichos sistemas pueden adquirir que en muchas ocasiones la información que finalmente se publica en los portales está compuesta por estadísticas elaboradas previo procesamiento de los datos obtenidos, pero en pocas ocasiones se publican los datos reales, lo que provoca una disminución de la usabilidad de estos. RTOD (*Real Time Open Data*) surge como plataforma de intercambio de información entre aplicaciones institucionales no accesibles públicamente y en las cuales se alojan los datos en tiempo real, y el propio repositorio de datos abiertos, permitiendo así la posibilidad de automatizar la publicación de estos datos.

El funcionamiento se basa en el uso de diferentes drivers que comunican el proveedor de contenidos en tiempo real con el servidor RTOD. Para llevar a cabo las comunicaciones de forma exitosa se dispone de drivers para gestionar la información procedente de diferentes fuentes y formatos, entre los que se destaca la posibilidad de gestionar la información procedente de la plataforma de control de ciudades inteligentes Smart City FIWARE. [3] En la **Figura 2.4** se puede ver la arquitectura de dicha plataforma:

[3] https://www.ogoov.com/es/rtod/

Figura 2.4: Arquitectura RTOD. Fuente: ogoov.com



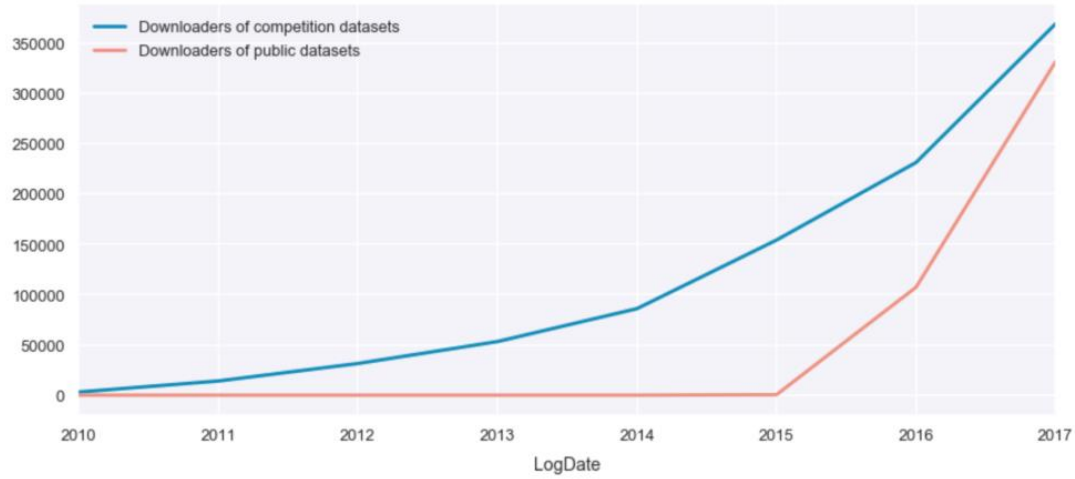
2.2. OTRAS INICIATIVAS

Además de las iniciativas *Open Data* existen otros portales y comunidades de usuarios en los que se comparten *datasets* de diversas temáticas para su descarga y uso de forma libre. Un ejemplo de estas comunidades es *Kaggle,* la mayor comunidad online de científicos de datos y machine learners la cual destaca por la organización de competiciones de *Machine Learning* en las que los usuarios participan de forma individual o en equipo con el objetivo de desarrollar el mejor algoritmo para resolver un problema propuesto por empresas de diferentes sectores. En estas competiciones, los usuarios parten en igualdad de condiciones disponiendo únicamente de la descripción del problema y los conjuntos de datos a analizar. A continuación, los participantes deben experimentar con diferentes modelos para encontrar la solución más precisa que será valorada a partir de métricas de evaluación basadas en un archivo de soluciones que se encuentra oculto a los participantes. Una de las ventajas más interesantes de las competiciones es que los usuarios tienen la posibilidad de compartir sus avances en la plataforma colaborativa *Kaggle Kernels*. A lo largo de la competición, las soluciones aportadas son evaluadas constantemente por los organizadores, mostrándose a los usuarios en un ranking realizado en función de la precisión del algoritmo aportado, el cual es comparado con un fichero con la solución del problema que permanece oculta para los participantes. Una vez se llega a la fecha límite, los ganadores reciben una recompensa económica a cambio de la solución propuesta de la que sale beneficiada la empresa.

En cuanto a la evolución de esta plataforma, desde su creación en el año 2010 y durante sus primeros 5 años de vida fue conocida principalmente por la organización de las competiciones anteriormente citadas, en las cuales se manejaban datos únicamente accesibles para los participantes. No obstante, sería a partir del año 2016 cuando la plataforma apostaría por la apertura de ciertos conjuntos de datos en un portal específico sin necesidad de enfocar su utilización al caso de las competiciones. Tal ha sido el éxito de este sitio que en 2017 el número de usuarios que descargaban datos públicos se situaba en 339.000, acercándose año a año al número de usuarios que

utilizan las descargas de *datasets* para competiciones (más de 350.000 en 2017). En el **Figura 2.5** se muestra una comparativa entre los usuarios que utilizan la plataforma como competición frente a los que la usa como repositorio de descargas.

Figura 2.5: Comparativa de número de usuarios de la plataforma pública de Kaggle frente a la plataforma Kernels de competición. Fuente: Blog Kaggle, 2017



2.3. INDICADORES DE CALIDAD DE DATOS ABIERTOS

Dada la variedad de tipos de datos que existen en los portales de datos abiertos y las diferentes estructuras que estos presentan, surge la necesidad de establecer indicadores para evaluar la calidad del dato ofrecida por cada portal. Los tres criterios más utilizados para dicha valoración son la accesibilidad, el uso y la reutilización de los datos. En el informe COTEC (Abella et al., 2017) se pueden ver los resultados de un estudio sobre el análisis de 103 portales de datos abiertos en España a partir de 6 criterios de evaluación: disponibilidad de mecanismos de publicación de las actualizaciones de datos; disponibilidad de un catálogo de recursos, número de juegos de datos disponibles y si el catálogo es descargable; existencia de mecanismos de conexión directa con los datos (API); disponibilidad de un portal donde se identifiquen servicios basados en los datos de los portales y número de servicios identificados; utilización de un sistema de gestión de datos; y el número de juegos de datos publicados. Por otro lado, en el informe, también resulta interesante la aplicación de una métrica para analizar el grado de madurez de portales de datos considerando la difusión, usabilidad y la reutilización de los datos de la siguiente manera:

- Difusión de los datos: tener más de 30 juegos de datos (peso 20%).

- Usabilidad del portal: Tener una fuente con actualizaciones del catálogo (peso 10%) y utilizar un sistema de gestión de datos (peso 15%).

- Reutilización de los datos:
    - Disponibilidad de un interfaz de programación de aplicaciones (API) para interacción automatizada con los juegos de datos (peso 25%).
    - Portal de aplicaciones / servicios basados en datos abiertos (peso 30%).

Uno de los criterios más conocidos para valorar la reutilización es el modelo presentado por Berners-Lee (2006) en el que se puntúa la calidad del portal con una valoración entre una y cinco estrellas en función de diferentes indicadores:

Tabla 2.1: Clasificación del nivel de reutilización de datos abiertos de Berners Lee. Elaboración propia a partir de Abella et al. (2018)

Valoración	Tipos de datos	Formatos disponibles
★	Datos y documentos no estructurados y accesibles desde licencias abiertas. Son poco manejables en el análisis y en muchas ocasiones difíciles de extraer.	Documentos pdf, imágenes en todos sus formatos
★★	Todo lo anterior además de datos o documentos estructurados en formato no abierto. Se requiere software propietario para su uso.	Todo lo anterior además de hojas de cálculo Excel
★★★	Todo lo anterior además de datos o documentos en formato estrucurado abierto.	Todo lo anterior además de hojas de cálculo en formato CSV (datos separados por limitador)
★★★★	Todo lo anterior además de datos referenciados mediante URIs	Todo lo anterior además de metadatos y formatos RDF.
★★★★★	Datos vinculados, contextualizados y disponibles en múltiples formatos	Todo lo anterior además de formatos linked data: RDF, URI, HTTP…

Además de este esquema de valoración por estrellas meramente descriptivo existen otros métodos más sofisticados para la evaluación del nivel de reutilización de los datos abiertos. Un ejemplo es la métrica *Meloda* (*Metric for Releasing Open Data*) que surge en el año 2011 cuando surge la necesidad de evaluar el nivel de reutilización de la información en una época en la que la calidad de los *datasets* en los portales *Open Data* era muy dispar. Aunque esta métrica fuera inicialmente creada para evaluar la calidad de los *datasets* y no de la totalidad de un portal, actualmente es utilizada para ambas finalidades, de modo que en ocasiones se utiliza para evaluar un gran número de conjuntos de datos de un mismo portal con el fin de aproximar el nivel global de reutilización de este. Tras la primera publicación, en sus primeras versiones se tomaba como referencia el cumplimiento de las tres siguientes dimensiones (Abella et al., 2014):

- *Estándares técnicos*: evalúa el uso de formatos abiertos en la información disponible en el portal.

- *Accesibilidad a la información*: mide el grado en que la información disponible en los *datasets* y documentos es completa y detallada, así como fácilmente utilizable por el usuario haciendo uso del software adecuado.

- *Dimensión legal*: la información estará abierta a los usuarios de acuerdo a un marco legal disponible en el portal en el que se especificarán la declaración responsable y los términos de reutilización basados en licencias que garanticen la libre reutilización de los datos.

Dado que la métrica se encuentra en constante evolución en consonancia con la aparición de nuevos portales y nuevas formas de mostrar la información, la última versión publicada en abril de 2017 (*Meloda v4.13*) añade tres nuevos criterios,

estableciendo así seis dimensiones de análisis. Además de las dimensiones descritas en las primeras versiones, la versión del 2017 añade:

- *Accesibilidad al modelo de datos*: analiza en qué medida el publicante comparte el modelo de datos y el seguimiento de estándares nacionales e internacionales. Por ejemplo: descripción de los campos que componen el dataset y demás información relevante que ayude al usuario a comprender el significado de los datos.

- *Información georreferenciada*: evalúa el grado en el que la información contiene ciertos campos con datos que aporten información adicional sobre su ubicación (dirección, ciudad, país, coordenadas…)

- Información en tiempo real: para este indicador se evalúa el grado en que los datos se encentran actualizados. Dependiendo del dataset, en ocasiones es aceptable que los datos sean actualizados anualmente, en cambio, en otros conjuntos de datos, por ejemplo, los procedentes de sensores en ciudades inteligentes, lo ideal es que los datos sean actualizados en periodos de tiempo muy cortos.

La métrica establece tablas de ponderaciones para cada uno de los indicadores en función de diferentes niveles de calidad en cada caso. En la **Tabla 2.2**, se resumen dichas puntuaciones:

Tabla 2.2: Puntuaciones de los indicadores de la métrica Meloda. Adaptación a partir de documentación de Meloda 4.13 (2017)[4]

Estándar	%	Acceso	%	Legal	%	Modelo de datos	%	Información georreferenciada	%	Tiempo real [5]	%
Estándar cerrado no reutilizable (pdf, imágenes…)	10	No accesible de forma automática vía web. Necesarios permisos de terceros)	0	Con licencia de Copyright	0	No dispone de modelo publicado. (ej: tablas de datos sin descripción de campos)	15	Sin información geográfica No se localiza ningún campo con datos sobre la ubicación de los datos.	15	Plazo de actualización de los datos de semanas.	15
Estándar cerrado reutilizable (shp) y estándares abiertos no reutilizables (Excel con macros)	35	Se requiere registro en la web y hay limitaciones en el número de accesos o descargas	10	Permite reutilización Sólo para uso privado	10	Dispone de modelo con campos de datos. (ej: tablas de datos con descripción de campos)	35	Campo de texto simple con una sola variable que aporta información de la localización de los datos (ej: ciudad)	30	Plazo de actualización de los datos de días.	49
Estándar abierto reutilizable (Excel, csv, txt…)	60	Acceso web a través de una URL única.	50	Permite reutilización sin fines comerciales	25	Dispone de modelos de datos con especificaciones tales como un vocabulario, en este caso no normalizado	50	Varios campos de georreferenciación (ej: ciudad, calle, país…)	50	Plazo de actualización de los datos de horas.	70
Estándar abierto con metadatos	100	Acceso web a través de	90	Permite reutilización	90	Modelo de datos local aceptado	90	Incluye localización precisa a través	90	Plazo de actualización de	90

[4] http://www.meloda.org/full-description-of-meloda/

[5] En caso de periodos de actualización variables, la métrica indica que la puntuación se calcularía como la media aritmética de todos ellos.

(rdf, json, xml…)		una URL con parámetros.		con fines comerciales		por organismos locales de estandariza-ción.		de coordenadas (latitud, longitud)		los datos de minutos.	
-	-	Incluye una sección para consultas SPARQL además de una API document ada.	100	Permite reutilización sin ningún tipo de restricciones o solo de atribución(citar propietario)	100	Dispone de un modelo global y aceptado por entidades globales de estandariza-ción.	100	Georreferenciación completa (todo lo anterior)	100	Plazo de actualización de los datos de segundos.	100

Una vez evaluado el conjunto de datos y establecidas las puntuaciones para cada parámetro (P_i) de la tabla anterior, se utiliza la siguiente fórmula para calcular el valor de la métrica, comparando posteriormente el valor obtenido con los rangos de reutilización explicados en la **Tabla 2.3**.

$$Meloda = \sqrt[6]{\prod_{1}^{6} P_i}$$

Tabla 2.3: Rangos para la valoración de la métrica Meloda. Adaptación de Meloda 4.13 (2017)

Rango obtenido	Valoración
0 - 25	Nivel de reutilización insuficiente
25 - 30	Nivel de reutilización básico
50 - 75	Nivel de reutilización avanzado con características mejorables
75 - 100	Nivel de reutilización óptimo

2.4. EVALUACIÓN DE LA CALIDAD DE LOS DATOS DEL PORTAL DE DATOS ABIERTOS DE SANTANDER

En el caso práctico que se propone en esta sección se utilizarán *datasets* pertenecientes al portal de datos abiertos del Ayuntamiento de Santander. En este capítulo se realizará una evaluación previa de la calidad del portal valorando tres factores clave como son la accesibilidad, la usabilidad y el nivel de reutilización de La información pública contenida en el catálogo de datos. Para ello, se hará uso de los modelos de evaluación explicados en el apartado 2.3.

2.4.1. Evaluación de la accesibilidad

El portal de datos abiertos de Santander dispone de una página informativa sobre las medidas adoptadas para la implantación políticas de accesibilidad en el propio portal. En este caso, se destaca el uso de tecnologías estándar X(HTML) + CSS3, las cuales permiten adaptar la página a cualquier navegador. Además, en el caso de los posibles contenidos multimedia que pudieran contener ciertas páginas, dichos elementos disponen de información adicional en forma de texto que se mostraría en el caso de haber problemas en cargar el contenido. Además, se ofrece la opción de descarga de plugins gratuitos como alternativa para solventar el problema.

Por otro lado, se señala el seguimiento de los siguientes estándares W3C:

- W3C XHTML 1.0
- W3C CSS3
- W3C WCAG nivel AA para el seguimiento de pautas de accesibilidad al contenido web.

A pesar de seguir una estructura intuitiva y con un alto grado de accesibilidad, destaca la ausencia de elementos y *plugins* que doten al portal de mayor accesibilidad para usuarios con ciertas discapacidades. Por ejemplo: botones para aumentar y disminuir el tamaño de fuente, elementos de audio para mostrar ciertos contenidos…

2.4.2. Evaluación de la usabilidad

La usabilidad de un portal de datos abiertos o una página web en general mide la facilidad con la que el usuario puede interactuar con ella. En el caso del portal de datos abiertos de Santander destaca positivamente la buena estructuración del portal, la limpieza del diseño distribuyendo los contenidos de los conjuntos de datos en diferentes secciones (en este caso, los llamados catálogos disponibles en la **Figura 2.4**), la búsqueda de la buena experiencia del usuario evitando ventanas emergentes y los diferentes componentes que permiten interactuar con el portal tales como las barras de búsqueda o los botones de valoración de los conjuntos de datos.

Como aspectos negativos se señalan la ausencia de un mapa web y la mala adaptación del contenido del portal a teléfonos móviles (*non-responsive website*) la cual tampoco se ve suplida por la presencia de una app.

2.4.3. Evaluación del nivel de reutilización de los datos

En este apartado se evaluará el nivel de reutilización de los datos haciendo uso de la métrica Meloda definida anteriormente. Dado el elevado número de conjuntos de datos disponibles, para realizar la evaluación global de reutilización del portal se evaluarán un total de doce muestras (2 conjuntos de datos por cada uno de los 6 catálogos en las que se clasifican estos).

Figura 2.4: Catálogos de datos disponibles en el portal Santander Datos Abiertos [Fuente: datos.santander.es]

Una vez evaluado cada uno de los indicadores en todos los conjuntos de datos seleccionados, se calculará su valor medio, el cual será utilizado para calcular la métrica. Las puntuaciones obtenidas se muestran en la **Tabla 2.4**:

Tabla 2.4: Cálculo de los parámetros de la métrica Meloda. [Fuente: elaboración propia a partir de datos de datos.santander.es]

Catálogo	Dataset	Estándar	Acceso	Legal	Modelo datos	Información georreferenciada	Tiempo real
Transporte	Histórico Semanal de Mediciones de Tráfico	60	100	100	35	15	49
	Aparcamientos Subterráneos	100	100	100	100	100	15
Urbanismo e Infraestructuras	Distritos y Secciones	100	100	100	100	50	15
	Parques y Jardines	100	100	100	100	90	15
Cultura y ocio	Monumentos	100	100	100	100	90	15
	Agenda Cultural	100	100	100	100	15	49
Medio ambiente	Sensores Ambientales	100	100	100	100	90	100
	Analítica del Agua	100	100	100	100	30	15
Ciencia y tecnología	Sensores de Parking de Superficie	100	100	25	100	90	100
	Sensores Móviles	100	100	100	100	90	100
Sociedad y bienestar	Noticias de Servicios Sociales	100	100	100	100	15	100
	Pulso de la Ciudad	100	100	100	100	15	100
VALOR MEDIO		96,67	100	93,75	94,58	57,5	56,08

Como se puede observar en los resultados de la **Tabla 2.4**, se trata de un portal en el que predomina el uso de múltiples formatos en los conjuntos de datos por lo que en general se sigue un estándar abierto con presencia de metadatos.

En cuanto a las URL de acceso se caracterizan por contener parámetros que identifican y relacionan la información a través de id's. Además, el portal en su conjunto dispone de una API para la realización de consultas SPARQL, por lo que en este aspecto el portal recibe la máxima valoración.

En el marco legal destaca que la casi la totalidad de los *datasets* analizados están sujetos a una licencia *Creative Commons de Atribución 4.0 Internacional* en la cual se permite al usuario la libertad de compartir y adaptar los datos bajo restricciones de atribución. Como excepción encontramos el conjunto de datos de sensores móviles de la ciudad el cual dispone de una licencia de *Atribución No Comercial 4.0 Internacional*.

En cuanto a la calidad del modelo de datos, destacan aquellos conjuntos que contienen información en formato RDF, la cual contiene vocabulario estandarizado en formato DCAT (Data Catalog) en el caso del portal de Santander. Por otro lado, aquellos catálogos cuyos conjuntos solamente están disponibles en formatos XLSX o CSV, presentan una valoración menor en el indicador de modelo de datos de la métrica

Meloda. Además, el portal dispone de una página con vocabulario[6] referido a aquellos *datasets* que reciben datos en tiempo real y que se identifican con el modelo Real Time *Open Data* descrito anteriormente, así como vocabulario relacionado con el callejero de Santander. En ambos casos se utilizan formatos Turtle y RDF/XML.

Por último, en cuanto al periodo de actualización de los *datasets*, predominan aquellos que se actualizan diariamente, destacando además los datos procedentes de ciertos sensores (de tráfico y ambientales principalmente) que son actualizados en tiempo real.

Con dicha información se procede a calcular el valor de la métrica:

$$Meloda = \sqrt[6]{96{,}87 * 100 * \ 93{,}75 * 94{,}58 * 57{,}5 * 56{,}08} = \ 80{,}737\%$$

Con el resultado obtenido y de acuerdo a los rangos proporcionados en la **Tabla 2.3**, se podría clasificar este portal con un nivel de reutilización óptimo. Del mismo modo, de acuerdo con la métrica de Berners-Lee (**Tabla 2.1**), se califica el nivel de reutilización del portal con 5 estrellas.

[6] http://def.santander.es/

CAPÍTULO 3. MARCO TEÓRICO DEL BUSINESS INTELLIGENCE

En este capítulo se va a realizar una revisión del estado del arte en lo que se refiere a las diferentes herramientas y estrategias de inteligencia de negocio que hoy en día pueden llevar a cabo las empresas partiendo de la evaluación del grado de madurez de implantación de este tipo de herramientas. Para ello, se estudiarán las fases de las que consta un proyecto de BI según diferentes metodologías teniendo en cuenta los principales enfoques tradicionales, así como los nuevos enfoques ágiles.

3.1. BUSINESS INTELLIGENCE: OBJETIVOS Y RELEVANCIA

Los primeros sistemas de información tenían como principal finalidad recopilar información para ayudar en la toma de decisiones. Sin embargo, estos han ido evolucionando a sistemas cada vez más informatizados cuyo objetivo es dar soporte a los procesos que se llevan a cabo en las organizaciones (producción, logística, contabilidad, ventas, gestión de recursos…) a través de los llamados sistemas de información de gestión (ERPs, CRMs...). Cando estos sistemas, además de ser útiles en la gestión empresarial, sirven de ayuda en la toma de decisiones, se habla de "sistemas de información para la toma de decisiones" y es en este punto donde la Inteligencia de Negocio juega un papel fundamental.

Definimos Inteligencia de Negocio como "*el conjunto de estrategias enfocadas a la administración y creación de conocimiento sobre el medio, a través del análisis de los datos existentes en una organización o empresa*" (Pérez, 2015).

Para la extracción de dicho conocimiento, se utilizan diversas herramientas basadas en sistemas de información para el tratamiento de los datos de la organización en cualquiera de sus ámbitos (económico, producción…). Para ello, es habitual el uso de procesos ETL (extraer, transformar y cargar) con el objetivo de filtrar y adecuar los datos de las diversas fuentes a las necesidades del analista y siguiendo las reglas de negocio.

Dichas herramientas y metodologías del BI comparten las siguientes características:

- *Accesibilidad*: garantizan al usuario el acceso a la información independientemente de la fuente de procedencia.

- *Apoyo en la toma de decisiones*: no deben limitarse a la mera muestra de información en diferentes formatos, sino que deben permitir al usuario realizar un análisis más profundo de los datos.

- *Orientación al usuario final*: el objetivo de los sistemas de información más actuales es ser lo suficientemente intuitivos como para ser utilizados por cualquier usuario con unos conocimientos mínimos, pero sin necesidad de una formación exhaustiva.

3.2. TIPOS DE PROYECTOS DE BUSINESS INTELLIGENCE SEGÚN SU GRADO DE MADUREZ

Existe una gran variedad de técnicas, herramientas y metodologías aplicadas en el ámbito del análisis de datos. En este apartado se pretende realizar un breve estudio con el que aportar una visión general sobre estas, definiendo los fundamentos teóricos en los que se sustentan, así como aportando ejemplos de aplicaciones prácticas de cada herramienta.

González (2014) hace mención del modelo de madurez de BI propuesto por Wayne Eckerson, el que fuera director de TDWI (*The Data Warehousing Institute*), y cuya representación se muestra en la **Figura 3.1**. Este modelo, basado en la Integración de Modelos de Madurez y Capacidades (CMMI), fue concebido inicialmente para evaluar la madurez del uso de almacenes de datos en las empresas siendo posteriormente adoptado por TDWI para darle un uso más genérico en la evaluación del grado de implantación de procesos y herramientas de BI. Según el modelo, descrito a partir de una campana de Gauss, la mayor parte de las empresas se encuentran en la tercera y cuarta etapa de un total de seis que se describen a continuación:

Figura 3.1: Modelo de Madurez de la Inteligencia de Negocio. Fuente: CantabriaTIC

En las dos primeras etapas: prenatal e infancia, la compañía solamente hace uso de informes obtenidos directamente de los sistemas de producción haciendo uso habitualmente de libros de Excel o ficheros de texto plano. En este nivel de madurez, el personal de TI es el máximo responsable de las herramientas de Inteligencia de Negocio y los informes solamente llegan a la alta dirección de la empresa y responsables de ciertas áreas. Durante la etapa de infancia, la organización hace uso de varias fuentes de datos (*Spreadmarts*) en diferentes formatos.

En la tercera etapa, denominada niñez, se comienzan a crear *Data Marts,* que representan divisiones individuales del almacén de datos especializadas para cada departamento. Además, se mejora la accesibilidad de los datos dando permisos a ciertos usuarios.

En la adolescencia, la compañía ha implementado *Data Marts* en todas las áreas de negocio y departamentos, por lo que surge la necesidad de unificar y estructurar dicha información en un almacén de datos común. En esta etapa el compromiso es mayor ya que la empresa dispone de personal propio o subcontratado con conocimientos en BI, que se encargan de implementar nuevas herramientas y cuadros de mando en los que un mayor número de usuarios pueden acceder al análisis de indicadores clave de rendimiento o KPI's cuya misión es medir el desempeño de los diferentes procesos de la empresa directamente ligados al cumplimiento de objetivos.

En la etapa adulta ya existe una solución implementada a nivel corporativo y se crea una administración MDM (*Master Data Management*). Este sistema no está únicamente implementado a nivel operacional, sino que forma parte de las decisiones estratégicas de la empresa, existiendo un alto compromiso por parte de todos los estamentos de la organización. En este punto se va un paso más allá del almacén de datos, el cual se convierte en un *Enterprise Data Warehouse* (EDW). La principal diferencia entre ambos

es que este último utiliza una base de datos unificada para toda la compañía y accesible a todos los departamentos (TECHNOPEDIA, 2018).

Por último, en la etapa de sabiduría, la compañía da el paso de crear un centro de competencia de BI certificado formado por un grupo de empleados encargados exclusivamente de llevar a cabo proyectos de este tipo en beneficio del resto de la empresa, permitiendo incluso el desarrollo de soluciones a medida. Además, los sistemas de información empresariales pasan a ser inter-empresariales por lo que se les proporciona acceso a proveedores y clientes con el objetivo de obtener beneficios adicionales a partir del análisis de los datos que estos puedan proporcionar.

En un informe publicado recientemente por la empresa tecnológica Efor (2018) se relacionan los distintos grados de madurez propuestos en el modelo de TDWI con los tipos de proyectos que se realizan en cada etapa y las tecnologías empleadas en cada caso. En la **Tabla 3.1** se muestra a modo de resumen dicha correspondencia:

Tabla 3.1: Tipos de proyectos BI en función de la etapa del modelo de madurez. Elaboración propia a partir de Efor: Los 5 grados de madurez de un proyecto BI.

Etapa Modelo TDWI	Implicación de la organización	Tipo de Proyecto BI y tecnologías empleadas
Prenatal	Solo personal IT	Procesos ETL + Reporting estático
Infancia	Solo peronal IT	Data Marts departamentales + Reporting
Niñez	A nivel departamental	Implantación de Data Warehouse
Adolescencia	A nivel departamental	Data Warehouse + OLAP
Adulto	A nivel de negocio	EDW + Cuadro de mando integral + Sistemas de soporte a las decisiones
Sabiduría	A nivel estratégico, grupo epecífico de BI	Sistemas interempresariales, Análisis de grandes datos, Data Mining,

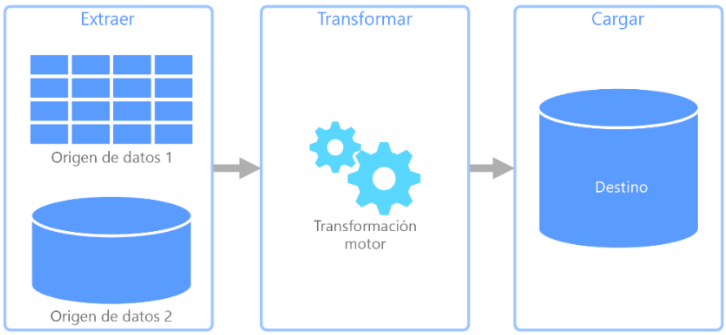
De acuerdo con la tabla anterior, se procede a explicar las diferentes tecnologías, procesos y metodologías utilizadas en cada caso de menor a mayor nivel de complejidad en los siguientes apartados.

3.2.1. Procesos ETL

Para la correcta realización de un proyecto de BI es necesario el seguimiento de una serie de fases en las que se profundizarán en las siguientes secciones. En primer lugar, es habitual hacer uso de los llamados procesos ETL para la adquisición, adecuación y almacenamiento de los datos que posteriormente se utilizarán las distintas herramientas de inteligencia de negocio.

ETL (*Extract, Transform and Load*) es la abreviatura del proceso que permite "*recopilar datos de varios orígenes, transformar los datos según las reglas de negocio y cargarlos en un almacén de datos de destino."* (Microsoft, 2017)

Figura 3.2: Proceso ETL. Fuente: Microsoft Azure Docs[7]



En concreto, en cada una de las tres etapas se realizan las siguientes operaciones:

- *Extracción*: consiste en la obtención de los datos procedentes de diferentes fuentes y formatos desde el sistema de origen. Habitualmente, estos datos son obtenidos desde bases de datos internas (sistemas OLTP, repositorios históricos internos…) o externas a la compañía, conocidas comúnmente como *bases de datos operacionales*. Se deberá de hacer un análisis previo de los datos para verificar si cumplen o no los requerimientos esperados. De esta forma, se pretende analizar la calidad del dato, su correcta estructuración y la seguridad del uso de estos con el objetivo de darlos por válidos o descartarlos.

- *Transformación*: se trata de integrar y adecuar los datos extraídos de acuerdo a las reglas de negocio. Se deberán manipular o convertir los datos con el objetivo de filtrar los datos más relevantes y hacer que éstos sigan un formato uniforme. Algunos ejemplos habituales de transformación de datos son la selección de columnas, la omisión de columnas con valores nulos, codificar valores, trasponer filas o columnas, dividir columnas, generar nuevas columnas a partir de cálculos con datos de otras etc.

- *Carga*: en esta última etapa se procede a cargar los datos ya transformados en el sistema de destino. En este punto, es de gran utilidad para la compañía disponer de un almacén de datos independiente de los sistemas de información utilizados diariamente para el desempeño del trabajo. Es importante en este punto saber diferenciar los sistemas basados en las bases de datos transaccionales OLTP (*On-Line Transactional Processing*) y los almacenes de datos (*Data Warehouse*) que es el lugar donde se almacenan los datos tras finalizar el proceso de carga.

3.2.2. Sistemas basados en OLTP, Data Warehouse y OLAP

Uno de los principales obstáculos con los que las compañías se enfrentan a la hora de llevar a cabo proyectos de BI es poder trabajar con los datos necesarios sin entorpecer las tareas habituales de los empleados y, por lo tanto, afectando en la menor medida posible a los sistemas de información con los que cuenta la empresa orientados a los procesos diarios como pueden ser la gestión de compras, ventas, logística, pedidos, gestión de relaciones con clientes etc. Dichos sistemas, están basados habitualmente

[7] https://docs.microsoft.com/es-es/azure/architecture/data-guide/relational-data/etl

en el procesamiento de transacciones en línea (OLTP). Para enfrentar esta problemática surgen los Almacenes de Datos o (*Data Warehouse*)

A continuación, se muestran las diferencias fundamentales entre ambos tipos de sistemas.

Tradicionalmente, se define un almacén de datos como "*una colección de datos orientados por tema, variables en el tiempo y no volátiles que se emplea como apoyo a la toma de decisiones estratégicas*" (Inmon, 1992).

Como se puede observar, en la cita anterior, la definición de los almacenes de datos se sustenta en cuatro características fundamentales:

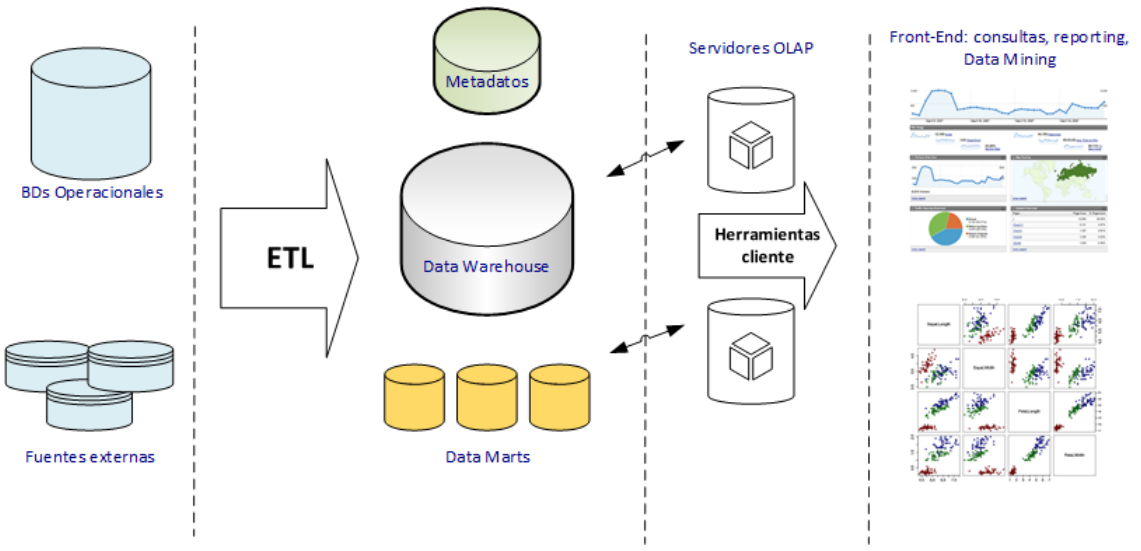
- *Orientación por tema*: los datos son organizados en diferentes bloques según las diferentes áreas y procesos de la empresa con el objetivo de mejorar la accesibilidad y agilizar las consultas. Dichas divisiones especializadas del almacén de datos son los llamados *Data Marts*.

- *Integración de los datos*: se debe garantizar la integridad de los datos a la hora de diseñar el almacén. Aspectos como evitar los nombres duplicados de elementos de las tablas de la base de datos o garantizar la integridad referencial del modelo, deben ser tratados en la etapa de transformación del proceso ETL.

- *Variación temporal y no volatilidad*: con el objetivo de preservar una base de datos que sirva como histórico, si un dato de las bases de datos operacionales es actualizado, los datos del DW no serán alterados, sino que se insertará un nuevo registro, manteniendo el anterior. De esta forma, el tamaño del DW se incrementa con el tiempo.

Con las propiedades anteriormente comentadas, queda definida la arquitectura más frecuentemente usada para definir un DW. Dicho esquema relaciona los distintos elementos y procesos involucrados en BI comentados anteriormente.

Figura 3.3: Arquitectura multi-nivel de un Data Warehouse. Fuente: elaboración propia



Esta estructura se asemeja a una arquitectura cliente-servidor que, en este caso, se compone de tres niveles o capas:

En el primer nivel se definen los procesos ETL y el almacenamiento de los datos en el almacén, que a su vez contiene *Data Marts* especializados para las distintas áreas de negocio y almacenes de metadatos. El DW hace las funciones de servidor de bases de datos, en este caso relacionales.

La capa intermedia contiene el servidor OLAP (*On-Line Analytcal Processing*) que tiene como fin optimizar el uso consultas para grandes volúmenes de datos minimizando los tiempos de respuesta. Se diferencian tres tipos de tecnologías OLAP:

La tecnología MOLAP (OLAP Multidimensional) se ayuda de los llamados cubos OLAP, que son bases de datos multidimensionales donde los datos se almacenan en un vector multidimensional. Este tipo de almacenamiento mejora la eficiencia de las consultas tipo SELECT en múltiples tablas, por lo que son más adecuados para la adquisición de datos por parte de software de BI. Por el contrario, las bases de datos OLTP suelen contener bases de datos relacionales ya que agilizan las consultas INSERT, UPDATE y DELETE.

Por otro lado, la arquitectura ROLAP (OLAP Relacional) permite la implementación de un modelo de datos multidimensional sobre bases de datos relacionales donde se ejecutan las consultas SQL.

Por último, la arquitectura HOLAP (OLAP Híbrido) combina características de las dos anteriores por lo que almacena ciertos datos en bases de datos relacionales y otros en multidimensionales.

Estas dos capas inferiores de la arquitectura se corresponden con el back-end de la arquitectura. Por último, existe una capa superior, que es la más cercana al usuario (front-end) y en la que se encuentran las aplicaciones de Inteligencia de Negocio utilizadas para la consulta de datos, elaboración de informes o Minería de Datos entre otras funcionalidades.

3.3. HERRAMIENTAS CLIENTE

En la arquitectura típica de los procesos de *Business Intelligence* que se detalla en la **Figura 3.3** se contemplan tres tipos de herramientas y técnicas básicas utilizadas por el usuario final para el análisis, visualización e interpretación de los datos. Estas son: herramientas EIS (*Eecutive Information System*), herramientas de informes y consultas avanzadas y herramientas y técnicas de Minería de Datos. En algunas literaturas (Pérez, 2015), se incluye un cuarto elemento: las herramientas OLAP. Sin embargo, estas se definirán más adelante ya que en muchos casos forman parte de los motores de consulta de los sistemas que se muestran a continuación:

3.3.1. Herramientas EIS

Los EIS (*Executive Information System*) también conocidos como ESS (*Executive Support System*) son software de apoyo orientado a la alta dirección de las compañías y que permite a éstos acceder a través de una interfaz gráfica sencilla a la información más relevante referente a la gestión de las diferentes áreas de la empresa. Estos sistemas se limitan a analizar los indicadores clave o KPI's y muestra los resultados en forma de gráficos y tablas con el fin de garantizar un análisis inmediato de los procesos más críticos. Disponen de un cuadro de mando integral para presentar los indicadores potenciales en el formato más adecuado además de alarmas automatizadas para alertar a la alta dirección en caso de que el valor de ciertas métricas no sea el adecuado. Esto sistemas de información suelen trabajar con almacenes de datos restringidos

específicamente para el personal ejecutivo y modelos multidimensionales de datos (cubos) para la explotación de la información.

3.3.2. Reporting y consultas avanzadas

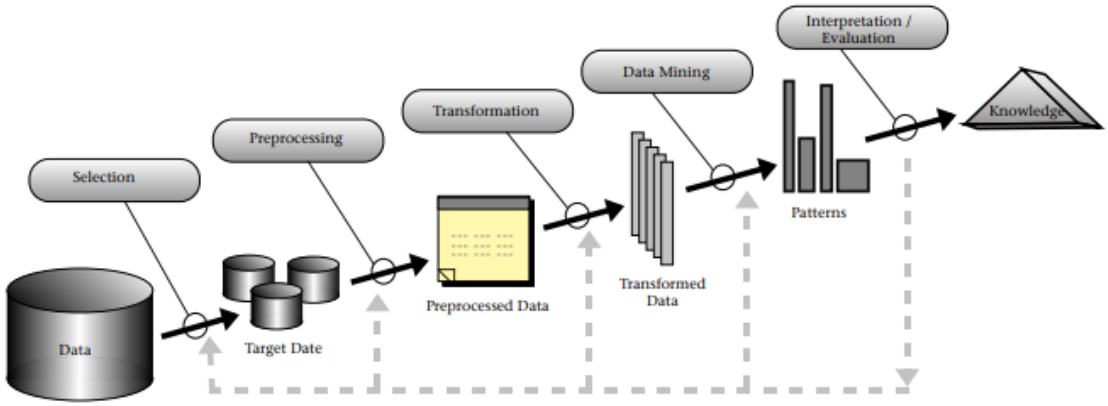
Las herramientas de consultas avanzadas e informes (*Query & Reporting*) se basan en sistemas relacionales que hacen uso de operadores clásicos como la selección, el agrupamiento, la concatenación etc. mostrándose el resultado en forma de tablas e informes. La principal diferencia entre los sistemas de consultas avanzadas y los de generación de informes reside en que los primeros están basados en una interfaz más sencilla sobre la que se permite una conexión a la base de datos y la ejecución de consultas a través de una consola en lenguajes tales como SQL, para posteriormente mostrar los resultados en tablas. Por su parte, las herramientas de reporting utilizan interfaces más elaboradas que permiten al usuario mapear la información en gráficos y tablas, lo que permite una experiencia más visual.

Para este tipo de sistemas la forma óptima de trabajo es la basada en los almacenes de datos, ya que permite agilizar el tiempo de respuesta de las consultas, así como simplificar la complejidad de esta.

3.3.3. Minería de Datos

SAS Institute define el concepto de *Data Mining* como el proceso de seleccionar, explorar, modificar, modelizar y valorar grandes cantidades de datos con el objetivo de descubrir patrones desconocidos que puedan ser utilizados como ventaja comparativa respecto a los competidores. La minería de datos es solamente una de las etapas del Proceso de Extracción de Conocimiento (*KDD*) propuesto por Fayyad et al. (1996), y que se muestra en la **Figura 3.4**. Este proceso contempla soluciones más avanzadas que las dos estrategias anteriores. Se utiliza principalmente para el reconocimiento de patrones que sigue un conjunto de datos, estimación de modelos, aproximaciones por regresión, tendencias etc. por lo que tiene una fuerte componente estadística, así como una componente informática en el uso de diferentes algoritmos.

Figura 3.4: Data Mining como parte del proceso KDD (Fayyad et al., 1996)



2.3.3.1. Metodologías de Minería de Datos

Para la elaboración de un proyecto de *Data Mining* existen diferentes metodologías a seguir que pueden ser distinguidas según su carácter *tradicional* o metodologías más abiertas y flexibles utilizadas frecuentemente en proyectos de ingeniería del software, las llamadas *metodologías ágiles*. En este apartado se hará un repaso a modo de comparación entre las diferentes metodologías existentes.

Metodologías tradicionales

Las metodologías tradicionales son aquellas que están documentadas a través de procedimientos muy detallados con el objetivo de evitar el riesgo y enfocadas al seguimiento del proyecto. (Mariscal et al. 2013). A continuación, se resumen dos de las metodologías tradicionales más utilizadas actualmente a nivel empresarial:

Se define la metodología SEMMA como el proceso de selección, exploración y modelado de grandes cantidades de datos para descubrir patrones de negocio desconocidos. El proceso consta de cinco fases, las cuáles se resumen a continuación:

En el primer paso del proceso, tras identificar el problema a resolver, se realiza un *muestreo* representativo de todos los datos disponibles. Es habitual que dicho muestreo o selección de datos se realice de forma aleatoria (*muestreo aleatorio simple*) para garantizar su validez. En esta fase también se pueden realizar particiones de los datos que son de gran utilidad a la hora de aplicar ciertos algoritmos supervisados que necesitan un conjunto de datos de entrenamiento, un conjunto de datos de test o un conjunto de datos para la validación del modelo.

El siguiente paso de la metodología SEMMA es la *exploración* de la información seleccionada con el objetivo de encontrar relaciones entre variables buscando comprender el significado de los datos a través de su visualización con la ayuda de gráficos tales como los histogramas de frecuencia, los diagramas de dispersión o incluso la aplicación de algoritmos de clustering. En esta etapa se busca detectar relaciones entre los datos, así como anomalías tales como la presencia de nulos (ausencia de datos) o la presencia de valores atípicos que se desvían significativamente del resto de los datos. El fin último de esta fase es seleccionar las variables más indicadas como entradas para el modelo de datos.

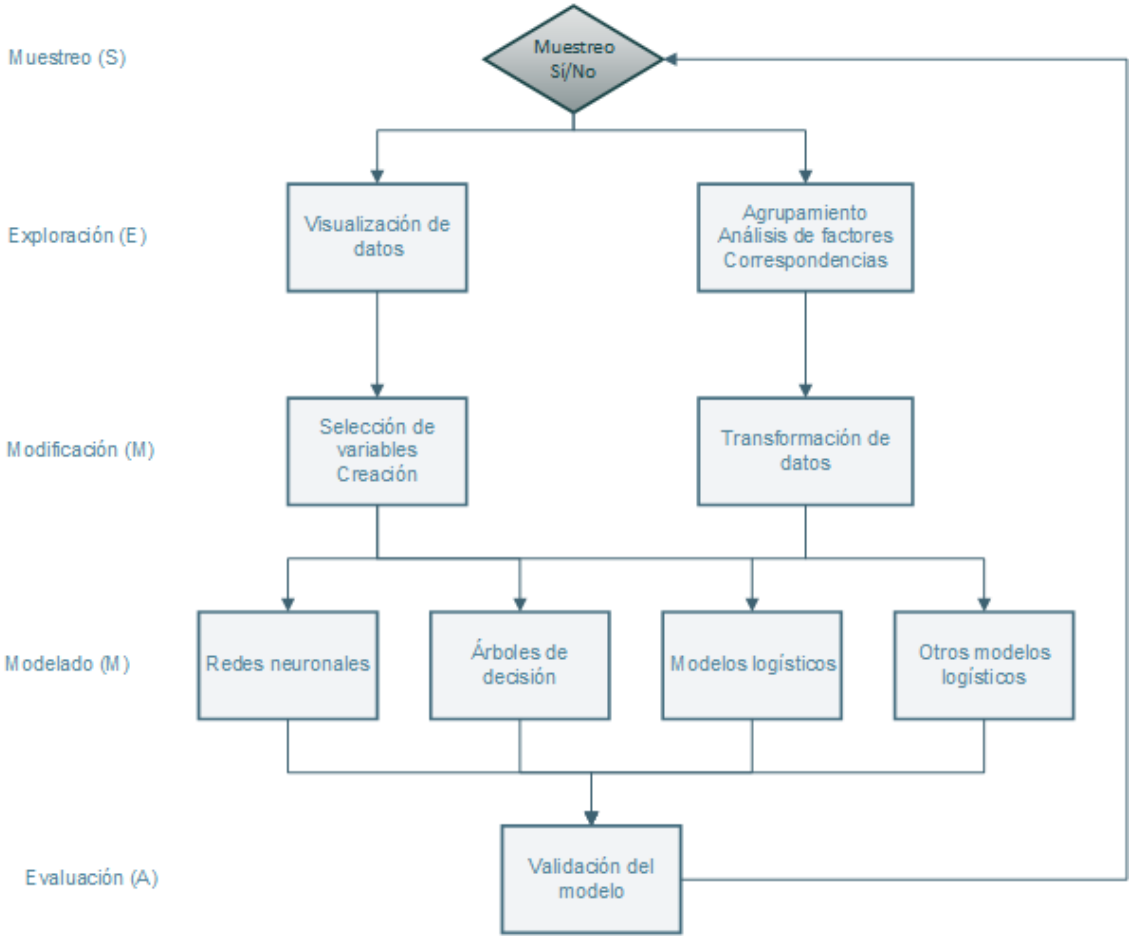
A continuación, en la etapa de *manipulación* se realiza una transformación de los datos de acuerdo a la información obtenida en la fase anterior con el fin de adecuarlos al formato requerido para aplicar sobre ellos las técnicas de *Data Mining*.

Una vez las entradas del modelo de datos poseen un formato adecuado para su uso se procede a la etapa de *modelado*. En esta fase se elabora el algoritmo, modelo o técnica a utilizar sobre las variables de entrada seleccionadas.

Una vez aplicado el modelo, se procede a la última etapa del proceso que es la *validación* de la precisión de dicho modelo y en qué grado se ajusta a los conjuntos de datos utilizados. Para ello, se utilizan métodos estadísticos que forman parte de las llamadas pruebas de bondad de ajuste.

Una de las principales características de esta metodología, es la retroalimentación entre diferentes etapas, de forma que en cualquier instante del proceso es posible regresar a la etapa anterior.

Figura 3.5: Etapas de la Minería de Datos según metodología SEMMA. Elaboración propia a partir del modelo de SAS Institute



Otra metodología comúnmente utilizada es **CRISP-DM** (*Cross-Industry Standard Process for Data Mining*) la cual fue elaborada conjuntamente por varias empresas en el año 2000 (Chapman et al., 2000). Esta metodología divide el ciclo de vida del proyecto en 6 fases, las cuales se componen de diversas tareas divididas en subniveles donde se explican los procedimientos a seguir en diferentes situaciones, organizándose en los primeros las tareas más generales y en los siguientes los procedimientos más específicos para llevar a cabo esas tareas. Esta metodología de carácter cíclico incluye la figura del cliente final dentro del proceso no dando por terminado el proyecto vez validado el modelo de datos, sino que contempla las labores de implementación, registro y documentación de cada una de las tareas y mantenimiento de la solución desarrollada. Al igual que ocurre en la metodología SEMMA, existe retroalimentación entre fases por lo que es posible regresar a etapas anteriores del proyecto en caso de que sea necesario.

A continuación, se resumen las 6 fases de CRISP-DM:

- *Fase I. Comprensión del negocio*: consiste en fijar los objetivos y el alcance del proyecto de acuerdo a los requerimientos del cliente. Esta fase incluye tareas como la recopilación de la información del negocio necesaria para la contextualización y el planteamiento del problema, la evaluación de riesgos, los

criterios de éxito, así como la elaboración de un primer plan de proyecto (provisional) de Minería de Datos.

- *Fase II. Comprensión de los datos*: en esta fase se trata de seleccionar y familiarizarse con los datos que posteriormente serán utilizados. Tras la tarea de recolección, se evalúa la calidad de los datos y se buscan relaciones entre éstos (tendencias hacia ciertos valores, agrupaciones…) que en ocasiones permiten obtener cierto conocimiento y formular las primeras hipótesis.

- *Fase III. Preparación de los datos*: en la tercera etapa se pretende transformar, filtrar, limpiar, integrar y dar un formato uniforme a los datos para su correcto uso a la hora de aplicar el posterior modelo de datos. Esta etapa es semejante a la tercera (Modificación) de la metodología SEMMA.

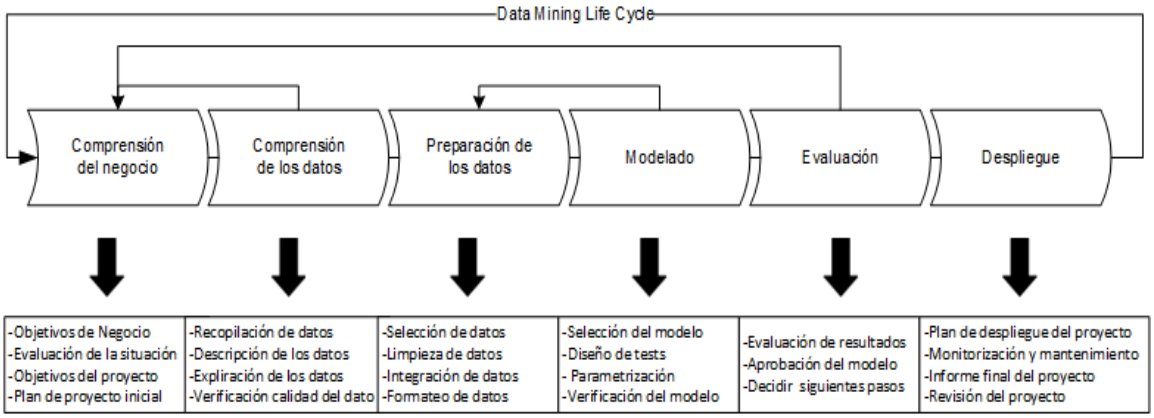
- *Fase IV. Modelado*: incluye tareas tales como la selección de las técnicas de Minería de Datos que se van a aplicar, el testeo del modelo de datos mediante el particionado de los *datasets* y la aplicación de algoritmos de validación, configuración de parámetros requeridos por el modelo y evaluación de la precisión del modelo de acuerdo a los resultados obtenidos tras aplicar los test. Es habitual tener que regresar a la etapa de preparación de los datos cada vez que se aplica un nuevo modelo debido a los diferentes requisitos de cada uno de ellos.

- *Fase V. Evaluación*: en esta etapa se evalúa el modelo desde la perspectiva de los criterios de éxito del negocio que previamente habían sido fijados en los informes generados al finalizar la primera fase del proyecto. Por esta razón, existe una retroalimentación en el modelo entre esta fase y la primera. Por otro lado, antes de implantar la solución desarrollada, se revisarán los procesos llevados a cabo en las etapas anteriores en busca de posibles aspectos a mejorar.

- *Fase VI. Despliegue*: en la última fase de esta metodología se elabora el plan de proyecto definitivo, que incluye además un plan específico de monitorización y mantenimiento que se lleva a cabo una vez implementada la solución.

En la **Figura 3.6** se muestra el ciclo de vida de CRISP-DM y las diferentes tansiciones de un estado a otro del proyecto.

Figura 3.6: Ciclo de vida del proyecto de Minería de Datos según metodología CRISP-DM. Elaboración propia a partir del modelo de Nicole Leaper (2009)



Metodologías ágiles

(Mariscal y otros, 2013) proponen varias referencias en las que se presentan analogías entre las metodologías ágiles diseñadas para ser aplicadas en el ámbito de la Ingeniería del Software y metodologías ágiles para llevar a cabo proyectos de *Data Mining*. Estas metodologías buscan un seguimiento del proyecto menos estricto que las tradicionales, pero a la vez más flexible y enfocado a la toma de decisiones en el caso de cambios o incidencias que pudieran perturbar la planificación inicial del proyecto. Por lo tanto, se da una mayor importancia a las exigencias del cliente, del que se recibe más *feedback* a lo largo del proyecto, el cual se organiza de forma incremental y con un mayor número de entregas.

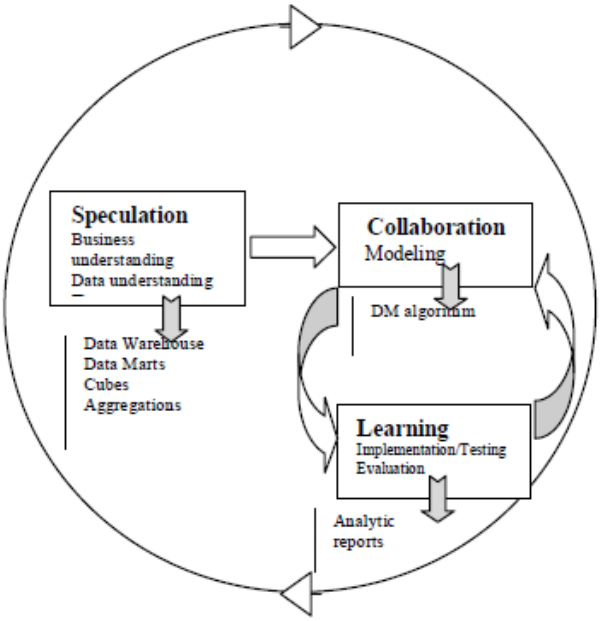
Varios autores han propuesto adaptaciones de metodologías ágiles ya existentes para su aplicación en proyectos de análisis de datos. No obstante, cabe destacar que estas metodologías están enfocadas a proyectos en los que se aplica una técnica específica de minería de datos, aunque se espera que en un futuro próximo se produzca una evolución metodologías ágiles genéricas orientadas a la elaboración de proyectos de análisis y minería de datos independientemente de la técnica aplicada en éstos. En el siguiente ejemplo, se muestra una metodología ágil aplicada a algoritmos de predicción.

Siguiendo esa dirección, Alnoukari y otros (2009) proponen la metodología ASD-DM (*Adaptative Software Development for Data Mining*) aplicada en soluciones de BI. Estas metodologías adaptativas son especialmente útiles en entornos en los que las reglas de negocio cambian constantemente. Para ello, proponen un caso práctico en el que se plantea la implantación de un *Data Mart* destinado al almacenamiento de datos de un servicio de atención al cliente siguiendo el modelo adaptado que se muestra en la **Figura 3.7**, el cual combina la metodología ASD original con las etapas que caracterizan a un proyecto de *Business Intelligence*.

En este caso, el ciclo de vida del proyecto consta de tres fases: en una primera fase de *especulación* en la que se emplean una gran cantidad de recursos del proyecto y donde se incluyen las etapas de comprensión del negocio, procesos ETL y creación del almacén de datos, *data marts* y cubos. A continuación, en la fase de *colaboración* se selecciona el modelo o algoritmo de predicción del que se va a hacer uso. En esta fase se requiere un alto grado de comunicación entre los miembros del equipo y

stakeholders. Por último, la etapa de *aprendizaje* engloba las tareas de testeo y verificación de los algoritmos empleados. En este punto del proyecto, las partes implicadas debatirán si el modelo propuesto cumple las expectativas para, en caso afirmativo proceder al despliegue de la solución y la realización de un informe de los resultados. En caso contrario, se regresaría a la fase de colaboración con el fin de seleccionar un algoritmo más adecuado. Como se observa en la **Figura 3.7**, se trata de un modelo cíclico dada la naturaleza cambiante del negocio, por lo que es una metodología abierta a modificaciones en cualquier instante del proyecto.

Figura 3.7: Modelo ASD-DM como adaptación de la metodología ASD para Minería de Datos. Fuente: Alnoukari et al. (2009)



En conclusión, se están adoptando metodologías ágiles propias de la ingeniería del software con el objetivo de adaptarlas a proyectos de Minería de Datos. Las principales ventajas de estas son la flexibilidad y la retroalimentación constante en las diferentes fases del proyecto, no obstante, Mariscal y otros (2013) proponen ciertas desventajas a tener en cuenta a la hora de utilizar estas metodologías. Entre ellas destacan la poca efectividad en proyectos de gran magnitud o proyectos críticos en los que se deben establecer todos los requisitos al principio del proyecto, evitando así la aparición de nuevos requisitos volátiles que retrasarían en gran medida la finalización a tiempo de las correspondientes tareas.

2.3.3.2. Clasificación de técnicas de Data Mining

El término *Data Mining* puede estar referido a múltiples actividades con distintas finalidades. (Berry & Linoff, 2000) proponen una clasificación para diferenciar las diferentes técnicas empleadas en la etapa de modelado de datos. A continuación, se agrupan dichas técnicas en función de si utilizan algoritmos de aprendizaje supervisado o algoritmos de aprendizaje no supervisado.

Técnicas de aprendizaje supervisado

- *Clasificación*: consiste en examinar las características de un conjunto de datos u objetos y asignar cada uno de ellos a una clase predefinida o etiqueta. Los objetos a clasificar suelen estar representados por registros en una base de datos y pueden tener formatos muy variados, texto plano, imágenes, señales de comunicaciones… Para ello, es habitual dividir los datos del *dataset* en dos conjuntos: un primer conjunto de datos de entrenamiento formado por elementos previamente clasificados y un segundo conjunto cuya clase es desconocida. El objetivo es utilizar el primer conjunto para elaborar un modelo que permita clasificar con la mayor precisión posible el segundo conjunto.

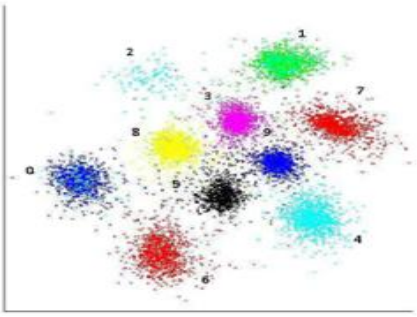
Un ejemplo sencillo de aplicación se puede encontrar sobre el *dataset* MNIST que almacena un vector de 10.000 imágenes (muestras) con trazos de números manuscritos entre 0 y 9. Esta base de datos está preparada para trabajar sobre ella omitiendo las etapas de pre-procesado de los datos, siendo el objetivo del problema utilizar un algoritmo que aprenda a averiguar qué número se ha escrito en cada caso y evaluar las prestaciones de dicho algoritmo en función del número de aciertos.

Figura 3.8: Captura del *dataset* MNIST. Fuente: Kaggle.com

Para la resolución de este problema, se puede recurrir a algoritmos de *Machine Learning* como el KNN (*K-Nearest Neighbours*). Este algoritmo supervisado se basa en el concepto de distancia euclídea para realizar la clasificación de las muestras. Por otro lado, existen otros modelos que pueden ser utilizados en problemas de clasificación de mayor complejidad tales como las redes neuronales, los árboles de decisión o las Máquinas de Vectores de Soporte (SVM).

En la **Figura 3.9** se ha aplicado el algoritmo de clasificación DNet-KNN (*Deep Neural Network KNN*) sobre el dataset MNIST descrito anteriormente. Como salida se obtienen las muestras clasificadas gráficamente mediante un conjunto de agrupaciones o *clusters*, lo que ayuda a entender mejor el algoritmo. Por ejemplo, nótese que las nubes de puntos correspondientes a los elementos clasificados como unos y sietes se encuentran cercanas en la gráfica. Esto es debido a que los números de forma manuscrita tienen cierta similitud (menor distancia euclídea entre ellos), al igual que ocurre con las muestras correspondientes a los números 8 y 5. Otro aspecto a tener en cuenta es que las muestras más alejadas del centro (centroide) de cada nube de puntos son las que más probabilidades tienen de ser asociadas a una clase errónea y, en consecuencia, reducir la precisión del clasificador.

Figura 3.9: Representación bidimensional de la clasificación de muestras del dataset MNIST mediante el algoritmo DNet-KNN. Fuente: Researchgate.net



- *Análisis predictivo*: la principal diferencia entre el análisis predictivo y otras técnicas como la clasificación o la estimación radica en que estas últimas parten del aprendizaje de resultados anteriores que sirven para entrenar el algoritmo, existiendo cierta incertidumbre en los resultados que se mide a través de la precisión del algoritmo empleado. En el lado opuesto se encuentran las técnicas de predicción, que se basan en estimaciones futuras del comportamiento de las variables involucradas en el problema. En este caso, la única forma de medir la precisión de la técnica empleada es esperar a comparar los resultados con lo que ocurra en el futuro.
  Una de las técnicas más empleadas para llevar a cabo análisis predictivo es la *regresión*. Esta técnica da como resultado un modelo basado en una función que ajusta un gran número de puntos que representan la población a estudiar. Este modelo puede ajustar la nube de puntos a diferentes tipos de funciones (lineal, segmentada por intervalos, parabólica...)

Técnicas de aprendizaje no supervisado

- *Estimación*: utiliza métodos estadísticos para estimar una variable aleatoria no observable $\boldsymbol{Y}$ a partir de otra variable aleatoria conocida $\boldsymbol{X}$ que debe estar relacionada con la anterior, de forma que se obtiene una estimación $\widehat{\boldsymbol{Y}}$ con el menor error posible, siendo habitual minimizar el error cuadrático medio MSE. Dichas estimaciones se pueden obtener a partir de diferentes funciones como pueden ser una función constante $\boldsymbol{\alpha}$, una combinación lineal $\boldsymbol{\alpha X + b}$ o bien una estimación sin restricciones que puede dar lugar a cualquier otro tipo de función.

- *Patrones de afinidad y asociación*[8]: en esta tipología de problema se relacionan dos o más ítems pertenecientes a un conjunto $I = \{i_0\ i_1 \ldots\ i_n\}$ de una misma transacción con el objetivo de predecir un patrón o regla que los relacione. Por ejemplo, definida la siguiente regla:

  $$\{i_0\ i_1\} \Rightarrow \{i_2\}$$

  que representa que la aparición conjunta de los elementos i_0 e i_1 implica la aparición del elemento i_2 en alguna de las transacciones de la base de datos. Para que estos patrones puedan ser considerados válidos, existen una serie de restricciones o

[8] R. Agrawal; T. Imielinski; A. Swami. 1993. Mining Association Rules Between Sets of Items in Large Databases", SIGMOD Conference, pp. 207-216

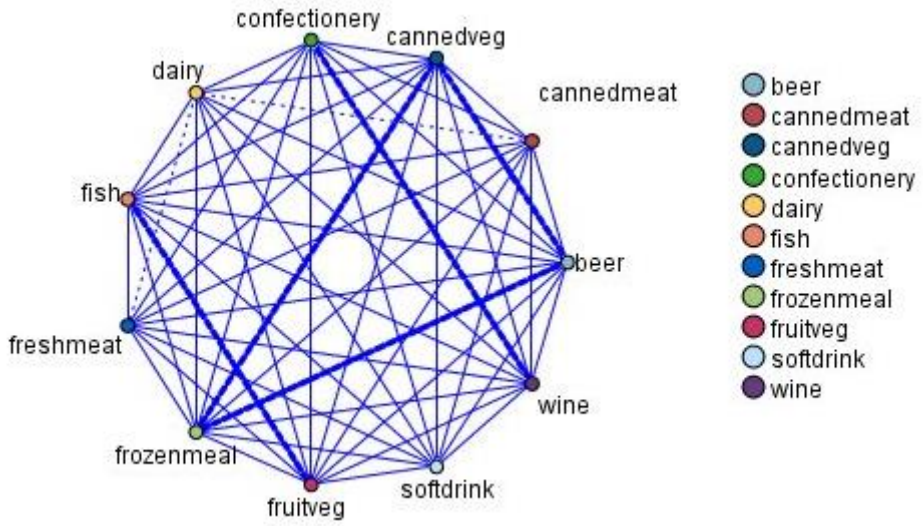
umbrales mínimos tales como los definidos en los conceptos de soporte y confianza, que sirven como métricas del grado de afinidad en las reglas. Se define soporte como el número total de apariciones de un conjunto de ítems en proporción con el número total de transacciones posibles en la base de datos $\boldsymbol{D}$. Por su parte, la probabilidad de encontrar un conjunto de elementos $\boldsymbol{Y}$ en una transacción sabiendo que el conjunto $\boldsymbol{X}$ pertenece a la transacción, representa el concepto de confianza de la regla. Matemáticamente, los conceptos anteriormente mencionados se definen como:

$$sop(\boldsymbol{X}) = \frac{|\boldsymbol{X}|}{|\boldsymbol{D}|}$$

$$conf(\boldsymbol{X} \Rightarrow \boldsymbol{Y}) = \frac{sop(\boldsymbol{X} \cap \boldsymbol{Y})}{sop(\boldsymbol{X})} = \frac{|\boldsymbol{X} \cap \boldsymbol{Y}|}{|\boldsymbol{X}|}$$

Un caso de uso frecuente de las reglas de asociación es la búsqueda de patrones de consumo en la cesta de la compra del consumidor, de forma que se busca concluir que la adquisición de un cierto producto implica frecuentemente la compra de otros, que pueden ser o no complementarios. Por ejemplo, de acuerdo con la red mostrada en la **Figura 3.10**, para dicho caso uno de los patrones más repetidos es la compra de cerveza, comida enlatada y congelados en una misma compra, dado que hay un mayor número de aristas (transacciones) que unen dichos nodos (ítems).

Figura 3.10: Red de asociación mostrando patrones de consumo en la cesta de la compra.
Fuente: IBM Knowledge Center (2009)

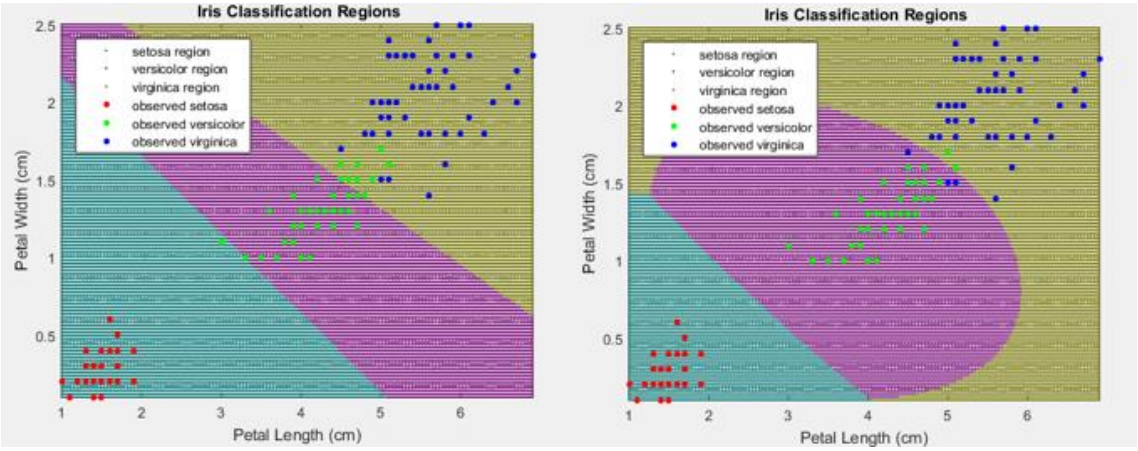


- *Clustering*: consiste en segmentar un grupo de datos de carácter diverso en subgrupos de datos que cumplan características similares (*clusters*). Para clasificar cada objeto en un grupo u otro se utiliza generalmente como criterio de referencia la distancia. En este caso no existen clases predefinidas ni datos de prueba para entrenar a los algoritmos, sino que tiene que ser el propio analista quien interprete los resultados obtenidos una vez aplicadas las técnicas de agrupamiento. El clustering tiene múltiples aplicaciones en campos tan diversos como la biología, la

meteorología, el marketing y los negocios. Por ejemplo, es una técnica comúnmente empleada a la hora de agrupar consumidores en distintos segmentos en función de los datos obtenidos en encuestas. De esta forma, se pueden observar las tendencias de consumo, los clientes potenciales y las relaciones entre diferentes tipos de clientes con el objetivo de desarrollar nuevos productos o entrar en nuevos mercados.

A continuación, se muestra un ejemplo sencillo en el que se aplican técnicas de clustering para agrupar muestras de pétalos de plantas de diferentes especias en función de las dimensiones de los pétalos. En este caso se han utilizado *Support Vector Machines* (SVM) con diferentes funciones kernel que definen la forma de las regiones que delimitan los diferentes hiperplanos de clasificación. De esta forma, se puede estimar a simple vista cuál es el modelo que mejor se ajusta a cada problema. Para este caso en concreto se ha utilizado un modelo lineal para la gráfica de la izquierda y un modelo polinómico para la de la derecha. Como se puede observar, el clúster para la especie *setosa* queda en ambos casos claramente definido dado que tiene unas características muy diferentes al resto de clases y en ambos modelos todas las muestras de esta especie se clasifican de forma correcta. Por otro lado, el primer modelo (lineal) clasifica de forma errónea 8 muestras, mientras que el polinómico reduce el error a 5 muestras, por lo que para este problema la segunda sería la opción más precisa.

Figura 3.11: Agrupación o clustering para la clasificación de muestras de pétalos (Fisher Irish dataset) aplicando SVM. Elaboración propia.



- *Análisis descriptivo*: permite entender las relaciones que pueden existir entre ciertas variables almacenadas en una base de datos de cierta complejidad. El análisis descriptivo va ligado en la mayor parte de los casos a diferentes técnicas de visualización de datos, siendo una de ellas el clustering, anteriormente comentada. Se trata de un análisis que no aporta un resultado concreto, sino que será el analista el que interprete las visualizaciones y obtenga, en consecuencia, sus propias conclusiones.

A modo de conclusión del capítulo, hay que tener en cuenta que las técnicas descritas en este apartado forman parte de las últimas etapas del proyecto. Éstas, junto a los procesos ETL y el uso de almacenes de datos se interrelacionan entre sí formando el esquema (ver **Figura 3.3**) que ayuda a entender las etapas de un proyecto de BI desde la adquisición de los datos, su almacenamiento en repositorios específicos, hasta el uso de dichas herramientas enfocadas al análisis de los datos.

CAPÍTULO 4. CASO DE APLICACIÓN DE TÉCNICAS DE BUSINESS INTELLIGENCE SOBRE DATOS ABIERTOS

41. ELECCIÓN DE DATASETS

Para la realización de este caso práctico se ha elegido el conjunto de datos del "Histórico Semanal de Mediciones de Tráfico"[9] del portal de datos abiertos de Santander. Este conjunto de datos recopila información de las mediciones de 258 sensores de tráfico situados en diferentes localizaciones de la ciudad de Santander. Los *datasets* se actualizan diariamente, recopilando información minuto a minuto del día anterior y mostrando en siete ficheros diferentes las mediciones para cada día de la semana. Estos datos son de especial interés para el Centro de Control de Tráfico Municipal para la regulación del tráfico y la programación de los semáforos. Cada dataset dispone de los siguientes campos:

- *Dc:identifier*:identificador único alfanumérico para cada conjunto de medidas en un instante de tiempo.

- *Ayto:medida*: identificador unívoco del sensor que, además permite identificarlo en los mapas representados a partir del dataset que contiene la ubicación de las espiras.

- *Ayto:intensidad*: número de vehículos contados a la hora.

- *Ayto:ocupación*: porcentaje de tiempo que la espira está ocupada por un vehículo.

- *Ayto:carga*: relaciona la intensidad de tráfico y la ocupación dando una idea del grado de congestión del tráfico.

- *Dc:modified*: indica la fecha y hora del conjunto de medidas realizadas. El intervalo horario de las mediciones se encuentra entre las 00:00 y las 17:00 horas.

Adicionalmente, se hará uso del recurso "Ubicación de Sensores", el cual almacena información geográfica sobre la ubicación de los sensores de espiras encargados de tomar las medidas del tráfico. Este conjunto de ficheros sigue el formato multiarchivo shapefile compuesto por archivos de diferentes extensiones que permiten relacionar varias bases de datos para representar los objetos sobre un mapa y etiquetarlos con atributos. Dichos archivos han sido convertidos en un único archivo en formato GeoJSON para facilitar el manejo de los datos.

[9] http://datos.santander.es/dataset/?id=historico-semanal-de-mediciones-de-trafico

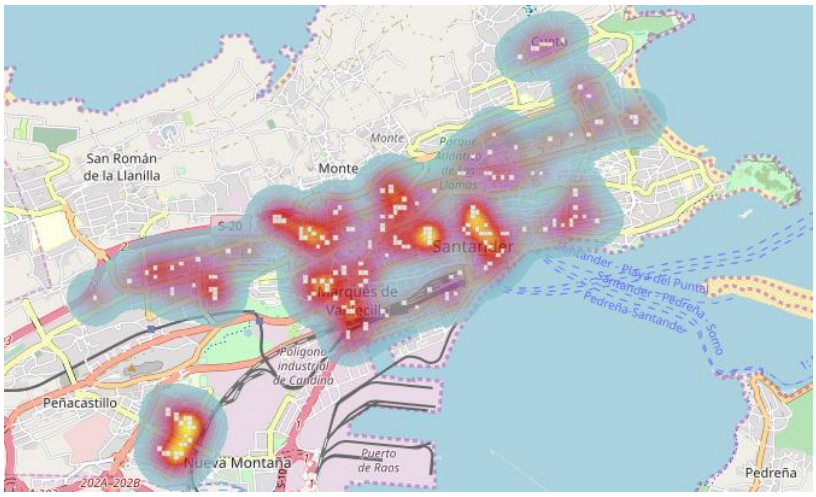
4.2. HERRAMIENTAS UTILIZADAS

Para este caso práctico se han elegido dos herramientas de análisis de datos: Microsoft Power BI para una primera fase de preparación y visualización de los datos y la herramienta de software libre Weka, la cual ofrece múltiples opciones para la aplicación de algoritmos de minería de datos. Además, se ha hecho uso de una plataforma en la nube para la representación de mapas A continuación, se describen brevemente las opciones más destacadas que ofrece cada una de ellas:

4.2.1. ArcGIS

Haciendo uso de la herramienta ArcGIS se ha obtenido un mapa en el que se representan los sensores de tráficos ubicados en las calles de Santander a partir de un archivo GeoJSON que permite representar la información de los sensores de forma georreferenciada, es decir, al pulsar sobre cualquier punto que represente un sensor en el mapa se pueden obtener sus coordenadas y el identificador del sensor. Por otro lado, la plataforma online de ArcGIS permite añadir capas al mapa sobre las cuales se pueden ejecutar diferentes análisis. Por ejemplo, en la **Figura 4.1** se muestra el análisis de mapa de calor, mediante el cual se pueden observar las zonas de la ciudad con mayor concentración de sensores. El motivo de la elección de ArcGIS es la posibilidad de mostrar mapas en PowerBI a través del componente “ArcGIS Maps for Power BI”, simplemente accediendo al perfil de la plataforma, lo que facilita en gran medida dichas visualizaciones.

Figura 4.1: Ubicación de los sensores y ejecución de análisis de mapa de calor a través de ArcGIS.



4.2.2. Microsoft Power BI

Power BI es una herramienta de análisis de datos proporcionada por Microsoft. Dispone de múltiples opciones para la importación de datos a través de diversas fuentes, incluyendo archivos en diferentes formatos (Excel, csv, xml, json…). Por otro lado, dispone de una interfaz de datos que permite visualizar estos en forma de tabla y crear nuevas tablas a partir de la transformación de los datos de origen. En la interfaz de informes se dispone de un lienzo sobre el que el usuario puede incluir diferentes gráficos para crear dashboards, así como añadir filtros a la información y obtener estadísticos. Por último, se incluye una sección para el modelado de datos, que permite relacionar tablas procedentes de múltiples fuentes que vayan a ser utilizadas en un mismo

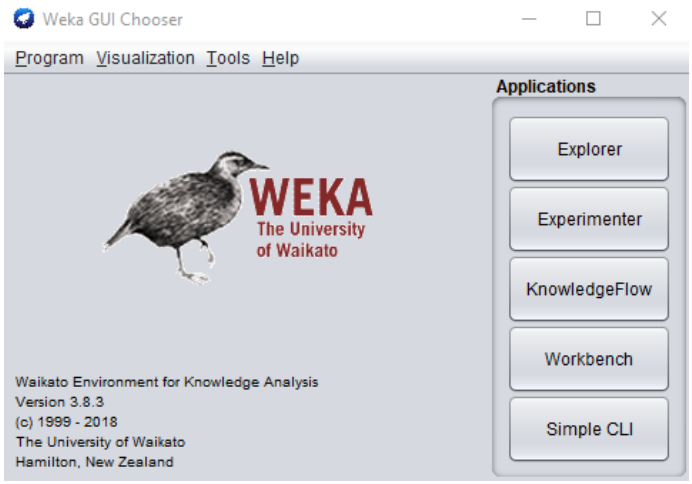
proyecto mediante un modelo relacional. En este trabajo se ha prestado especial atención a la transformación y visualización de los datos, ejecutando en ciertas ocasiones funciones del lenguaje de fórmulas DAX (Data Analysis Expressions) para la adaptación de los datos

4.2.3. Weka

Weka es un software libre y de código abierto que contiene diversos algoritmos de Machine Learning, herramientas de visualización y modelos predictivos implementados para la resolución de problemas de Minería de Datos. Las siglas Weka derivan de Waikato Environment for Knowledge Analysis, dado los orígenes de la aplicación, cuyo desarrollo comenzó en 1993 en la Universidad de Waikato.

El programa consta de una interfaz sencilla desarrollada en Java y se inicia con una GUI que permite escoger entre cinco aplicaciones:

Figura 4.2: GUI para la selección de aplicaciones de Weka



- *Explorer*: consta de la interfaz de pestañas que permite la carga y el pre-procesado y filtrado de los datos, la aplicación de diferentes algoritmos de Data Mining de clasificación, clustering y asociación, así como el uso de herramientas de visualización de la información.

- *Experimenter*: ofrece las mismas funcionalidades que el explorador, aunque destinadas a la ejecución de tareas a más largo plazo, dado que permite guardar la sesión actual para retomarla en otro momento. Uno de los principales usos de esta aplicación es la comparación de la precisión de diferentes algoritmos experimentando simultáneamente con ellos.

- *Knowledge Flow*: permite al usuario seleccionar distintos componentes y arrastrarlos a un canvas en el que puede diseñar distintos procesos o flujos en forma de diagrama, lo cual ayuda a comprender todo el proceso del análisis de datos de una forma más visual.

- *Workbench*: combina en una sola interfaz las diferentes interfaces de Weka, lo que mejora la experiencia del usuario al poder acceder a todas las aplicaciones desde una misma ventana.

- *Simple CLI*: permite acceder a todas las funcionalidades de Weka desde una interfaz de línea de comandos similar a la que se utilizaría para ejecutar código Java.

4.3. RESULTADOS OBTENIDOS

Para llevar a cabo este experimento se va a hacer uso de uno de los *datasets* del conjunto "Histórico Semanal de Mediciones de Tráfico" del portal *Open Data* Santander que ofrece datos en tiempo real. En primer lugar, se visualizarán los datos haciendo uso de la herramienta Power BI. Dado que el *dataset* elegido, correspondiente a la fecha 17-12-2018 contiene casi 475.000 registros, en una primera fase se obtendrán diferentes visualizaciones de las mediciones conjuntas en toda la ciudad, utilizando para ello diferentes tablas obtenidas a partir del conjunto original para facilitar la representación de los datos. No obstante, para la aplicación de ciertos algoritmos de Minería de Datos a través de la herramienta Weka, será necesario filtrar los datos con el fin de reducir su dimensionalidad para facilitar una ejecución más rápida de los algoritmos.

Para la obtención de los datos se ha utilizado la opción de importar de un archivo de "Texto o CSV", cargando desde una carpeta local el archivo csv con el *dataset* descargado previamente. Una vez hecho esto, se modificará la tabla de datos generada en Power BI para renombrar las columnas y transformar la columna *dc:modified*, la cual muestra en formato de texto la fecha y hora en una sola celda, en dos columnas separadas de fecha y hora. Esta modificación permitirá un acceso sencillo a los datos a la hora de representarlos. Para ello, se utilizarán las siguientes expresiones DAX, con las que se obtienen las dos nuevas columnas con el formato deseado:

```
Fecha = FORMAT(dia[Instante de Tiempo]; "DD/MM/YY")
Hora = FORMAT(dia[Instante de Tiempo]; "hh:mm")
```

En primer lugar, se han obtenido histogramas de frecuencia de la intensidad de tráfico, la ocupación y la carga para los datos obtenidos a lo largo del día señalado. Para ello, se han generado nuevas tablas en el modelo de datos en las que se ha hecho uso de fórmulas DAX para su creación. Para obtener una lista de los diferentes valores de cada parámetro sin repetición se ha utilizado las siguientes fórmulas, generando cada una de ellas una nueva tabla en la que se aplicará un flitro de orden ascendente:

```
Tabla_Carga = DISTINCT(dia[Carga])
Tabla_Ocupación = DISTINCT(dia[Ocupación])
```

Posteriormente, se añadirá una nueva columna a cada tabla en la que se realice un recuento del número de repeticiones de cada valor en el *dataset* original. De esta forma, se obtendrían los datos necesarios para construir los histogramas de frecuencia:

```
Frecuencia_Carga = COUNTX(FILTER(ALL(dia);dia[Carga]='Tabla_Carga'[Carga]);[Carga])
Frecuencia_Ocupación =
COUNTX(FILTER(ALL(dia);dia[Ocupación]='Tabla_Ocupación'[Ocupación]);[Ocupación])
Frecuencia_Intensidad =
COUNTX(FILTER(ALL(dia);dia[Intensidad]='Tabla_Intensidad'[Intensidad]);[Intensidad])
```

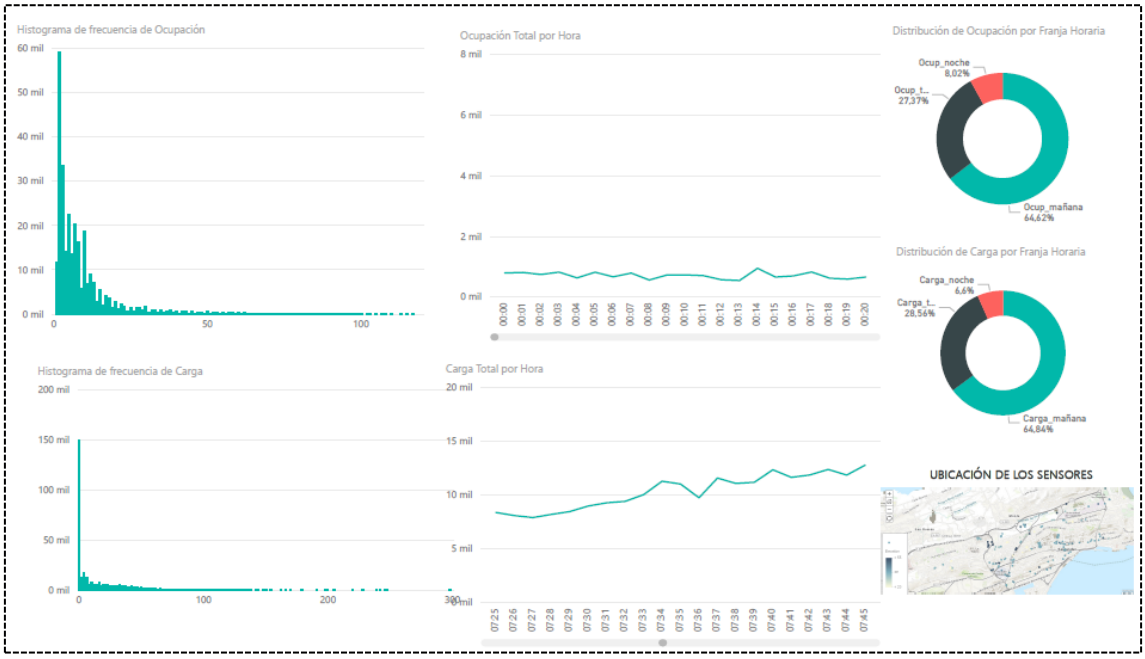
Una vez transformados los datos originales, se procede a obtener las diferentes representaciones gráficas que formarán parte de un dashboard. En este caso se han elegido las siguientes visualizaciones proporcionadas por Power BI Desktop:

- *Gráfico de columnas apiladas*: se utilizará para obtener histogramas de frecuencia de la intensidad, carga y ocupación a partir de las nuevas tablas generadas anteriormente para este propósito.
- *Gráfico de líneas*: mediante este gráfico se hará una representación temporal de la variación de cada uno de los parámetros a lo largo del día.

- *Gráfico de anillos*: se obtendrá la suma de cada uno de los parámetros en diferentes franjas horarias, dando así una idea de la congestión del tráfico por la mañana, tarde y noche. Para ello será necesario obtener medidas del total de cada parámetro en las diferentes franjas horarias creando una medida como la que se muestra a continuación para cada caso:

```
Ocup_mañana = CALCULATE(SUM(dia[Ocupación]);dia[Hora]>="07:00";dia[Hora]<"14:00")
```

Figura 4.3: Informe obtenido en Power BI



A continuación, se muestran los resultados obtenidos para la ocupación total:

Figura 4.4: Histograma de ocupación total

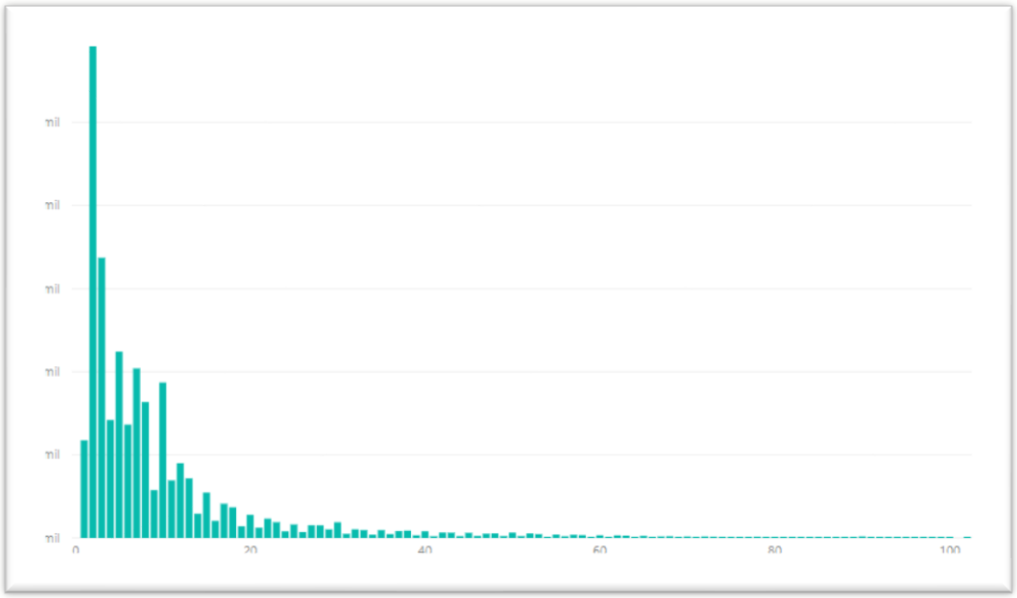
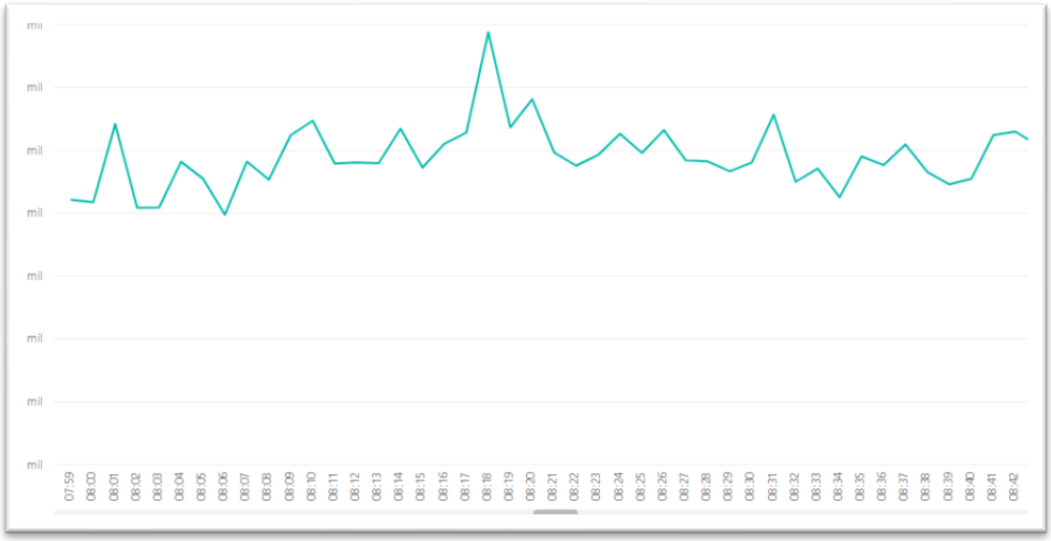


Figura 4.5: Variación temporal de la ocupación total



En vista a los resultados obtenidos tras analizar las medidas obtenidas durante un día, se observa una tendencia al aumento de tráfico en las primeras horas de la mañana y durante las tardes. Por el contrario, el tráfico disminuye considerablemente por las noches.

A continuación, se hace uso de la herramienta de análisis Weka con el objetivo de aplicar algunos de los algoritmos disponibles. Siguiendo la metodología propuesta por la aplicación, en primer lugar, se cargará el *dataset* escogido; en este caso contiene más de 400.000 instancias o filas por lo que será necesario aplicar un filtro para eliminar aquellas categorías (columnas) no relevantes a la hora de ejecutar los algoritmos, lo que facilitará en gran medida los tiempos de ejecución. En este caso, se suprime la primera columna en la que se muestran los identificadores de las medidas (*dc:identifier*)

Se plantea como objetivo la búsqueda de relaciones entre la ubicación de los sensores y las zonas con mayor influencia de tráfico. Para la resolución del problema, se han elegido algoritmos de aprendizaje no supervisado de clustering con el objetivo de hallar relaciones entre los identificadores de los sensores y las diferentes métricas de medición del tráfico.

Una vez interpretados los resultados y detectadas las relaciones en caso de que las hubiera, se procedería a cotejar los resultados con las ubicaciones de los sensores en el mapa obtenido en Power BI.

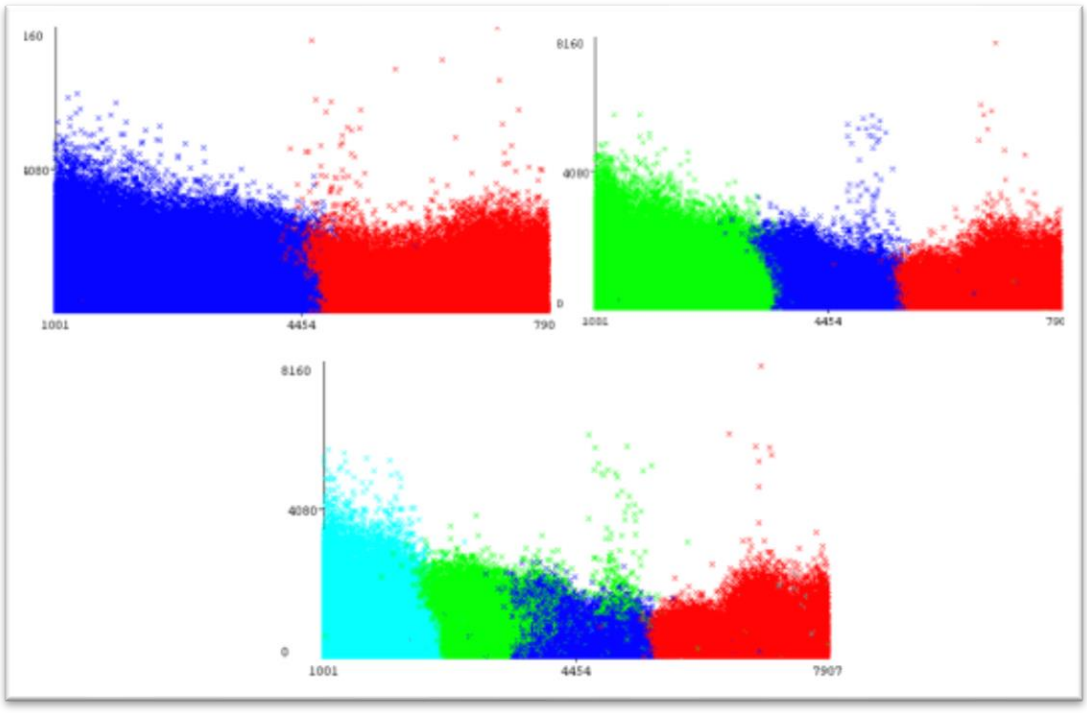
En la **Tabla 4.1** se mustran los resultados arrojados por el algoritmo K-Means ejecutado para la obtención de K=1, K=2 y K=3 clusters. Para cada caso se muestra el porcentaje de muestras clasificadas en cada agrupación así como la ubicación de los centroides en las visualizaciones obtenidas, que en este caso relacionan las ubicaciones de los sensores (atributo ayto:medida) en el eje de abscisas y la intensidad de tráfico en el de ordenadas.

Tabla 4.1: Resultados obtenidos de la aplicación del algoritmo K-Means

Algoritmo K-Means	K=2	K=3	K=4
Instancias por clúster	Clúster 0: 79% Clúster 1: 21% - -	Clúster 0: 24% Clúster 1: 16% Clúster 2: 60% -	Clúster 0: 20% Clúster 1: 16% Clúster 2: 34% Clúster 3: 31%
Centroides	Clúster 0: (3430, 324) Clúster 1: (2638, 366) - -	Clúster 0: (4116, 227) Clúster 1: (6810, 209) Clúster 2: (2275, 394) -	Clúster 0: (4208, 147) Clúster 1: (6810, 209) Clúster 2: (3037, 357) Clúster 3: (1658, 459)

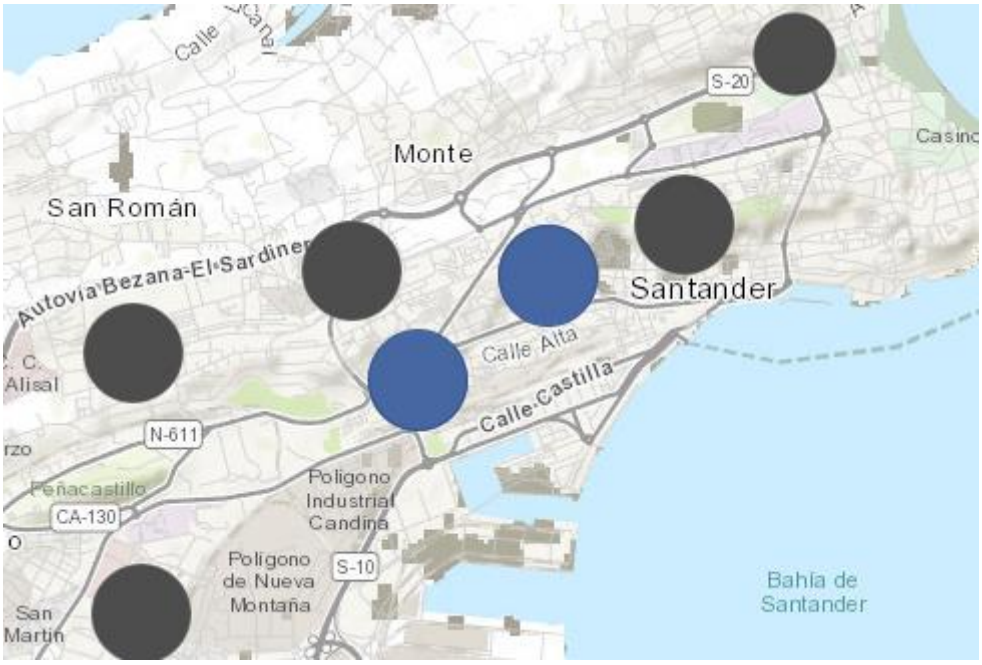
En la **Figura 4.6** se muestran los resultados obtenidos para cada una de las ejecuciones del algoritmo:

Figura 4.6: Aplicación del algoritmo K-Means con 2, 3 y 4 agrupaciones.



En vista a los resultados obtenidos en los dos primeros resultados en los que la precisión es mayor, los sensores que más intensidad de tráfico miden son aquellos identificados con los índices más bajos (entre 1001 y 2000). Si se aplica un filtro de agrupación sobre el mapa obtenido en ArcGIS, mostrando solamente dichos sensores se llega a la conclusión que las zonas pertenecientes a este grupo son las marcadas en azul en la **Figura 4.7**, que a su vez se corresponden a las zonas con mayor concentración de sensores, tal y como se mostraba en la **Figura 4.1**.

Figura 4.7: Agrupación y filtrado de las zonas con mayor intensidad de tráfico medida, de acuerdo a la interpretación del algoritmo de clustering.



Una vez obtenido el resultado, una métrica importante a evaluar en este tipo de algoritmos es la precisión. La herramienta Weka no ofrece una medida exacta de la precisión en el caso de los algoritmos de clustering. No obstante, la *métrica sum of squared errors* da una idea de qué valor de K agrupa mejor las muestras, obteniéndose en este caso mejores resultados para valores de K menores. En vista a los resultados obtenidos de forma gráfica se puede observar dicho aumento de errores, ya que a medida que se aumenta el número de clusters, la precisión disminuye al quedar cada vez menos definidas las nubes de puntos.

CAPÍTULO 5. CONCLUSIONES, LIMITACIONES Y LÍNEAS FUTURAS

Teniendo en cuenta los objetivos fijados en el primer capítulo, en este trabajo se ha mostrado la importancia que tiene para las AAPP no solo disponer de un portal de datos abiertos al ciudadano, sino también de preocuparse de que este siga unos criterios sólidos de accesibilidad, usabilidad y reutilización que permitan una buena experiencia del usuario.

Por otra parte, se ha pretendido mostrar ciertas herramientas que pueden ser de gran utilidad para analizar este tipo de datos y se han analizado diferentes metodologías y herramientas para la realización de proyectos relacionados con el análisis de datos, proponiendo un caso de uso en el que se han empleado herramientas de BI, Data Mining y visualización de mapas con el objetivo de sacar partido a grandes cantidades de datos. En este sentido, las principales dificultades encontradas han sido la necesidad de preparar los datos para obtener ciertos resultados (formateado de los datos, creación de nuevas tablas en el modelo a partir de fórmulas…). Asimismo, se han encontrado ciertas dificultades a la hora de representar el mapa de sensores a partir del archivo GEOJSON que contenía los datos de sus ubicaciones. Este problema ha sido suplido gracias a que en Power BI se localizó la posibilidad de sincronizar con mapas creados desde ArcGIS permitiendo un resultado más atractivo.

En cuanto al uso de la herramienta Weka, desde la que se aplicó el algoritmo de clustering K-Means, se han notado en ocasiones tiempos de ejecución muy altos. Para disminuirlos se recurrió a aplicar filtros que suprimiesen ciertos atributos irrelevantes para tal análisis como por ejemplo, el identificador de la medida.

Como líneas futuras se contempla explorar otras alternativas y métricas para la medición de la calidad de los datos y la evaluación de los portales *Open Data*, así como el uso de diferentes conjuntos de datos abiertos que permitan la aplicación de otros algoritmos.

BIBLIOGRAFÍA

- ABELLA A. [et al.]. 2014. Meloda, métrica para evaluar la reutilización de datos abiertos. *El profesional de la información,* **23**(6), pp. 582-588. ISSN: 1386-6710. Disponible en: https://recyt.fecyt.es/index.php/EPI/article/viewFile/epi.2014.nov.04/16932

- Abella, A.; Ortiz-De-Urbina, M.; De Pablos Heredero, C. 2017. La reutilización de datos abiertos: Una oportunidad para España. Informe COTEC (Fundación COTEC para la innovacción). Recuperado el 6-02-2019 de http://informecotec.es/media/INFORME_REUTILIZACION-DE-DATOS.pdf

- ABELLA A. [et al.]. 2018. Indicadores de calidad de datos abiertos: el caso del portal de datos abiertos de Barcelona. *El profesional de la información,* **27**(2), pp. 375-382. ISSN: 1699-2407. Disponible en: http://www.elprofesionaldelainformacion.com/contenidos/2018/mar/16.pdf

- AGENDA DIGITAL PARA ESPAÑA. 2015. *Plan Nacional de Ciudades Inteligentes*. [Consulta: 15 noviembre 2018]. Disponible en: http://www.agendadigital.gob.es/planes-actuaciones/Paginas/plan-nacional-ciudades-inteligentes.aspx

- ALNOUKARI, M. [et al.]. 2009. Applying ASD-DM Methodology on Business Intelligence Solutions: A Case Study on Building Customer Care Data Mart. En: *European Conference on Data Mining, ECDM: 18-20 junio 2009, Algarve. Portugal.* [Consulta: 12 septiembre 2018]. Disponible en: https://www.researchgate.net/publication/200450408

- BBVA API MARKET. 2016. *API REST: qué es y cuáles son sus ventajas en el desarrollo de proyectos*. [Consulta: 10 noviembre 2018]. Disponible en: https://bbvaopen4u.com/es/actualidad/api-rest-que-es-y-cuales-son-sus-ventajas-en-el-desarrollo-de-proyectos

- BERRY, M.; LINOFF, G. 2015. *Mastering Data Mining, The Art and Science of Customer Relationship Management.* Nueva York: Wiley. ISBN: 0-471-33123-6.

- BLOG KAGGLE. 2017. *Your Year on Kaggle: Most Memorable Community Stats From 2017.* 26 dic 2017. [Consulta: 19 septiembre 2018]. Disponible en: http://blog.kaggle.com/2017/12/26/your-year-on-kaggle-most-memorable-community-stats-from-2017/

- DATOS.GOB. 2011. *Proyecto Aporta*. 29 sep 2011. [Consulta: 20 agosto 2018]. Disponible en: http://datos.gob.es/es/documentacion/proyecto-aporta

- DATOS.GOB. 2018. *El OUR Data Index de la OCDE coloca a España en el sexto país en datos abiertos.* 01 ago 2017. [Consulta: 20 agosto 2018]. Disponible en: http://datos.gob.es/es/noticia/el-our-data-index-de-la-ocde-coloca-espana-en-el-sexto-pais-en-datos-abiertos

- DELOITTE. 2015. *Estudio y Guía metodológica sobre Ciudades Inteligentes*. [Consulta: 15 noviembre 2018]. Disponible en: https://www2.deloitte.com/content/dam/Deloitte/es/Documents/sector-publico/Deloitt_ES_Sector_Publico_Estudio-sobre-ciudades-inteligentes.pdf

- EFOR INTERNET Y TECNOLOGÍA. 2018. *Los cinco grados de madurez de un proyecto BI*. [Consulta: septiembre 2018]. Disponible en: https://www.efor.es/sites/default/files/madurez_de_los_procesos_de_bi.pdf

- FAYYAD, U. [et al.]. 1996. From Data Mining to Knowledge Discovery in Databases. *AAAI, Association for the Advancement of Artificial Intelligence*, **17**(3), pp. 37-54. ISSN: 0738-4602. Disponible en: https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131

- FUNDACIÓN CTIC. 2018. Public Dataset Catalogs Faceted Browser. [Consulta: 27 agosto 2018]. Disponible en: http://datos.fundacionctic.org/sandbox/catalog/faceted/

- GOBIERNO ABIERTO DE NAVARRA. 2018. Open Data y Reutilización de la Información del Sector Público. [Consulta: 25 agosto 2018]. Disponible en: http://www.gobiernoabierto.navarra.es/es/open-data/que-es-open-data/open-data-y-risp

- GONZÁLEZ. J.C. 2014: *¡Business Intelligence: El BI Maturity Model de Wayne Eckerson!*. *CantabriaTIC*, 10 de febrero. [Consulta: 14 noviembre 2018]. Disponible en: http://www.cantabriatic.com/business-intelligence-el-bi-maturity-model-de-wayne-eckerson/

- Leaper, N. 2009. A visual guide for CRISP-DM methodology. En*: EXDE: an experience design network*. 13 Mar 2009. [Consulta: 12 septiembre 2018]. Disponible en: https://exde.wordpress.com/2009/03/13/a-visual-guide-to-crisp-dm-methodology/

- LECUN, Y.; CORTES, C.; BURGES, C. 2013. *The MNIST database of handwritten digits*. 14 May 2013. [Consulta: 9 agosto 2018]. Disponible en: http://yann.lecun.com/exdb/mnist/

- MANJUNATH, M. 2018. *Designing and Implementing a Data Warehouse in the Cloud*. 29 Abr 2018. [Consulta: 10 julio 2018]. Disponible en: https://moodle.unican.es/pluginfile.php/379000/mod_resource/content/1/guia_citar_estilo_iso_1.pdf

- MARISCAL, G. [et al.]. 2013. *AgileDataMining: Un Enfoque Ágil para el desarrollo de proyectos de Data Mining*. Disponible en: https://www.researchgate.net/profile/Gonzalo_Mariscal/publication/234163466

- MELODA, PORTAL FOR DATA PUBLISHERS AND PROFESIONAL REUSERS OF DATA. 2017. *Full description of Meloda 4.13.* 2 abr 2017. [Consulta: 2 octubre 2018]. Disponible en: http://www.meloda.org/full-description-of-meloda/

- OECD ILABRARY. 2016. *Panorama de las administraciones públicas 2015*. 20 jul 2016. [Consulta: 22 agosto 2018]. Disponible en: https://www.oecd-ilibrary.org/governance/panorama-de-las-administraciones-publicas-2015_9789264262072-es

- OECD STATS. 2017. *Government at a Glance – 2017 edition: Open government*. [Consulta: 20 agosto 2018]. Disponible en: https://stats.oecd.org/Index.aspx?QueryId=78414

- OECD. 2017. *Government at a Glance*. 13 jul 2017. [Consulta: 20 agosto 2018]. Disponible en: http://www.oecd.org/gov/government-at-a-glance-22214399.htm

- PÉREZ LÓPEZ, C. 2007. *Minería de datos: técnicas y herramientas*. Madrid: Paraninfo. ISBN: 978-84-9732-492-2.

- PÉREZ MARQUÉS, M. 2015. *Business Intelligence. Técnicas, herramientas y aplicaciones*. Madrid: RC Libros. ISBN: 978-84-943055-2-8.

- SAS. 2018. *What is Data Mining?* [Consulta: 3 agosto 2018]. Disponible en: https://www.sas.com/en_us/insights/analytics/data-mining.html

- TECHNOPEDIA. 2018. *Enterprise Data Warehouse*. [Consulta: 12 agosto 2018]. Disponible en: https://www.techopedia.com/definition/26204/enterprise-data-warehouse

- W3.ORG. 2009. *Linked Data*. 18 jun 2009. [Consulta: 9 octubre 2018]. Disponible en: https://www.w3.org/DesignIssues/LinkedData.html

- WIKIPEDIA. 2018. OLAP. [Consulta: 15 julio 2018]. Disponible en: https://es.wikipedia.org/wiki/OLAP